

Researching Generative AI in Applied Linguistics

Edited by
Carol A. Chapelle,
Gulbahar H. Beckett, and Bethany E. Gray

IOWA STATE UNIVERSITY
Digital Press

Contents

Researching Language Learners’ GenAI Use 28
Researching GenAI in Language Assessment 31
Researching GenAI’s Capabilities 33
Does GenAI Work in Applied Linguistics? 34
References 36
Introduction 38
Methods 40
Participants 40
Data Collection 40
Data Analysis 42
Results and Discussion 42
RQ1: What are the Interaction Patterns of Participants with ChatGPT-driven Adaptive Feedback Pertaining to both Global and. 42
RQ2: To What Extent do Participants Exhibit Microgenesis within a Task and Subsequently Transfer that Knowledge across. 45
RQ3: What Reflections do Participants Provide regarding their Experiences with ChatGPT-driven Dialogic Assessment Sessio. 49
Conclusion 50
References 50
Appendix A 53
Appendix B 55
Appendix C 60
Appendix D 62
About the Authors 65
Introduction 67
Methods 69
Results 71
RQ1: How Effective and Precise are Students' Interactions and Communicative Skills in Formulating Response-seeking Quest. 71

Contents (continued)

RQ2: How cognitively demanding is the task of using ChatGPT, depending on learners' prior (meta)linguistic knowledge and. 73
RQ3: What are Students' Attitudes Toward ChatGPT’s Usability and Its Support for Developing Swedish and Academic Lang. 74
Discussion 76
Effectiveness and Precision of Interactions 76
Cognitive Demands 76
Usability and Perceived Learning Impact 77
Conclusion 77
Next Steps 78
References 79
About the Authors 80
Introduction 82
Literature Review 82
Methods 84
Context and Participants 84
Data Collection 85
Data Coding 86
Data Analysis 87
Results and Discussion 88
Nature of Interaction Patterns between Student and ChatGPT (RQ1) 88
Case-study Narratives: Potential for Learning (RQ2) 93
Learner Types in ChatGPT Interaction 96
Conclusion 97
References 99
Appendix A 102
About the Authors 105

Contents (continued)

Introduction	107
Methods	109
Study Setting and Design	109
Collection of Feedback Corpora	110
Data Analysis	110
Results	113
RQ1: How does Interacting with ChatGPT Influence the Linguistic Features and Evaluation Criteria that Students Use when G.	113
RQ2: To What Extent does ChatGPT Shape Students' Perspectives of Helpful Feedback?	122
Discussion	123
Conclusion	125
References	125
Appendix A	127
About the Author	128
Introduction	130
Methods	131
Project Context and Participants	131
Data Compilation	133
Analytical Approach	134
Results	134
Discussion	141
Conclusion	142
Acknowledgments	143
References	143
Appendix A	146
About the Authors	147

Contents (continued)

Introduction	149
Methods	152
Participants	153
Procedures	153
Results	154
Challenges in Humor Translation: Impact on Students' Ability to Convey Cultural Nuances and Achieve Dynamic Equivalence	154
GenAI’s Role and Prompt Engineering	159
Conclusion	160
References	161
Appendix A	163
Translation Activity 1: Una Optimista.	163
Translation Activity 2	166
Appendix B	167
About the Author	170
Introduction	172
Methods	175
Context and Participants	175
Human Tutoring and Feedback	175
SpeechAce: AI-Powered Speaking Evaluator	177
The Semi-Structured Interview	180
Data Collection Procedures	181
Data Analysis	181
Results	182
Learners’ Perceptions of Test Simulation Administrations	182
Learners’ Perceptions on Scoring	183

Contents (continued)

Learners’ Perceptions on the Feedback	185
Discussion	186
Conclusion	187
References	188
About the Author	191
Introduction	192
Methods	194
Participants and Contexts	194
Tasks	195
Rubric	195
Human Raters	196
ChatGPT-4 Ratings	197
Analysis	199
Results	200
Rater Agreement Across Criteria	200
MFRM	201
Discussion	208
Conclusion	211
Acknowledgments	211
References	212
Appendix A	214
Appendix B	218
Appendix C	219
Appendix D	220
Appendix E	221

Contents (continued)

About the Authors 222
Introduction 223
Methods 226
Research Context and Design 226
Participants 227
Instruments 227
Data Collection and Analysis 228
Results 230
Discussion 236
Conclusion 237
References 238
Appendix A 241
Appendix B 243
About the Authors 244
Author Note 244
Introduction 245
Study 1 246
Methods 247
Results 248
Discussion 253
Study 2: Responding to Multiple-Choice Items Without Reading Passages 254
Methods 255
Results 255
Discussion 257
Conclusion 258

Contents (continued)

References 258
About the Authors 260
Introduction 261
Methods 263
Results 266
Discussion 270
Conclusion 271
References 272
Appendix A 274
Appendix B 275
Appendix C 276
About the Author 277
Disclaimer 277
Introduction 279
Literature Review 279
Methods 281
Data 282
Coding Emotion 283
Coding Stance-Taking 284
Procedures 285
Results 286
RQ1: Emotions Expressed in Tweets 287
RQ2: Stance Expressed in Tweets 288
RQ3: Agreement Across Coders 290
RQ4: Differences between Human Coders and ChatGPT 292

Contents (continued)

Discussion 294
Conclusion 296
References 296
About the Authors 299
Introduction 301

Researching Generative AI in Applied Linguistics

Researching Generative AI in Applied Linguistics

Edited by

Carol A. Chapelle, Gulbahar H. Beckett, and Bethany E. Gray

Iowa State University Digital Press

Ames, Iowa

© 2025 Compilation and editorial content, Carol A. Chapelle, Gulbahar H. Beckett, and Bethany E. Gray.
© 2025 Individual chapters, the authors.

This work is published under a Creative Commons Attribution (CC BY) 4.0 International License.

The publisher is not responsible for the content of any third-party websites. URL links were active and accurate at time of publication.

Typeset in Times New Roman.
Book Cover Design: Hannah Litterer

ISBN: 978-1-958291-12-2
DOI: https://doi.org/10.31274/isudp.2025.211

Published by

Iowa State University Digital Press
701 Morrill Rd, Ames, Iowa 50011, United States

www.iastatedigitalpress.com

Iowa State University is located on the ancestral lands and territory of the Baxoje (bah-kho-dzhe), or Ioway Nation. The United States obtained the land from the Meskwaki and Sauk nations in the Treaty of 1842. We wish to recognize our obligations to this land and to the people who took care of it, as well as to the 17,000 Native people who live in Iowa today.

Contents

Editorial team *xiv*
Contributors *xv*
Acknowledgements *xvi-xvii*

Chapter 1. Introduction to Researching Generative AI in Applied Linguistics *1*
Carol A. Chapelle, Gulbahar H. Beckett, and Bethany E. Gray

Part I. Researching Language Learners' GenAI Use

Chapter 2. ChatGPT's Potential to Enhance Language Learners' Proximal Development through Dialogic Interaction *11*
Tuba Özturan and Prithvi N. Shrestha

Chapter 3. Leveraging ChatGPT for Language Learning: A Study on Foreign Teachers' Swedish Acquisition *40*
Sofie Johansson and Lina Larsson

Chapter 4. ChatGPT in ESL Writing: L2 Learners' Practices and Perspectives *55*
Inyoung Na, Mahdi Duris, and Volker Hegelheimer

Chapter 5. Investigating the Influence of ChatGPT Writing Guidance on Peer Review Assessments: An Exploratory Study *80*
Danilo Calle

Chapter 6. Empowering Youth through Language, AI, and Sustainable Advocacy: A Telecollaborative Approach *103*
Maria Mont, Melinda Dooly, and Onyemowo Angela Ejeme

Chapter 7. Unlocking the Potential of Generative AI: Enhancing Spanish Language Learning through Humor Translation Strategies *122*
Cristina Pardo-Ballester

Part II. Researching GenAI in Language Assessment

Chapter 8. Indonesian Learners' Perceptions on Human-AI Interaction in Speaking Assessments *145*
Andrias Susanto

Chapter 9. Comparing Human Raters' and ChatGPT-4's Scores in Measuring Content in CLIL Presentation *166*
Rie Koizumi, Saki Suemori, and Yusuke Kubo

Chapter 10. A Validation Study of AI-Generated Prompts in CILS (Certification of Italian as a Foreign Language) B2 Exams *197*
Giulia Peri, Sabrina Machetti, and Paola Masillo

Chapter 11. The Use of ChatGPT for Reading Test Development and Validation: Two Case Studies *219*
Hongli Li, Roula Aldib, and Chad Marchong

Part III. Researching GenAI's Capabilities

Chapter 12. Exploring ChatGPT's Ability to Recognize and Generate Language Variations: North vs. South Korean Dialects *235*
Jean Young Chun

Chapter 13. Comparing Human and Generative AI Analysis of Emotions in U.S. Abortion Tweets *252*
Sebnem Kurt and Mahdi Duris

Chapter 14. ChatGPT, Second Language Pragmatics, and Critical AI Literacy *274*
Robert Godwin-Jones

List of Tables

Table 2.1
ChatGPT-Generated Feedback Loop and Users' Reciprocity Acts

Table 2.2
Human-Generated Feedback Loop and Users' Reciprocity Acts

Table 2.3
Case 1 – Sue's Observed Errors Across Four Writing Sessions

Table 2.4
Case 2 – Ivy's Observed Errors Across Four Writing Sessions

Table 3.1
Percentage of Specific vs. General Prompts Learners Produced in Class and Out of Class.

Table 3.2
Usability and Learning Impact Ratings

Table 4.1
Characteristics of Participants

Table 4.2
Frequency of Prompt Types by Participant (n=8)

Table 4.3
Breakdown of 'Feedback (asking)' Types (n=8)

Table 4.4
Frequency of Feedback Methods Given by ChatGPT by Participant

Table 4.5
Frequency of Response Types by Participant

Table 4.6
Average Viewing Time per Feedback and Average Number of Turns across Sessions

Table 5.1
Student-Generated (SG) and ChatGPT-Generated (CG) Feedback Corpora

Table 5.2
Methods Used to Address the Research Questions

Table 5.3
Criteria for Coding Feedback Comment Types

Table 5.4
Normalized Mean Counts for Linguistic Features of Student-Generated Language

Table 5.5
Normalized Mean Counts for Linguistic Features ChatGPT-Generated Language

Table 5.6
Student-Generated and ChatGPT-Generated Linguistic Features Normalized Means

Table 5.7
Comparative Examples of Student-Generated and ChatGPT-Generated Language

Table 5.8
Friedman Statistical Test of Student-Generated Feedback

Table 5.9
Types of Feedback per Comment Indicating Evaluation Criteria Across Tasks

Table 5.10
Friedman Statistical Test of Differences Among Evaluation Criteria Indicated in Students' Feedback

Table 7.1
Cultural Joke Translation Using a Communicative Approach

Table 7.2
The Six Knowledge Resources of the GTVH for the Optimist Joke

Table 7.3
Adapting a Linguistic Joke to Create an Effective Equivalent Effect in Translation

Table 7.4
The Six Knowledge Resources of the GTVH for the Bus Joke

Table 8.1
Example Tutor's Notes Showing Typical Error Feedback that a Student Might Receive

Table 8.1
An Example of SpeechAce Score 7.1 out of 9.0 Interpreted on the CEFR and in Comparison to Standardized Tests (from https://www.SpeechAce.com/speaking-test/)

Table 9.1
Participants and Tasks

Table 9.2
Characteristics of Human Raters

Table 9.3
Four Datasets Analyzed

Table 9.4
Descriptive Statistics of All the Criteria Scores Combined

Table 9.5
Rater Agreement (Fleiss' Kappa [95% Confidence Interval])

Table 9.6
Global Model Fit and Facet Results

Table 9.7
Percentages of Variances Explained by Rasch Measures and Bias

Table 11.1
Comparison of Item Difficulty Estimates between CTT Empirical Results and AI Models

Table 11.2
Correlations among Item Difficulty Estimates across CTT Empirical Results and AI Models

Table 11.3
Comparison of Item Discrimination Estimates between CTT Empirical Results and AI Models

Table 11.4
Correlations among Item Discrimination Estimates across CTT Empirical Results and AI Models

Table 11.5
Summary of Item Responses

Table 12.1
Difference in Vocabulary between SK and NK

Table 12.2
Dialect Recognition Accuracy Rate by Term

Table 12.3
Multiple Logistic Regression Model Results for Passage Difficulty Levels and Dialects

Table 12.4
Average NK Word/Expression Identification Rate

Table 12.5
NK Word Meanings Provided by ChatGPT

Table 12.6
Types of NK Word Identification Errors

Table 12.7
Accuracy Rate of Conversion from SK Dialect to NK Dialect

Table 12.8
Linear Regression Model Results for Passage Difficulty Levels

Table 12.9
Types of Conversion Failures

Table 13.1
Results from Cohen's Kappa Analysis (N = 100 Tweets)

Table 13.2
Krippendorff's Alpha Results (N = 100 Tweets)

Table 13.3
Cohen's Kappa Analysis for Stance (N=100 Tweets)

Table 13.4
Chi-Square Contingency Table for Emotion Analysis

Table 13.5
Chi-Square Contingency Tables for Stance Analysis

List of Figures

Figure 2.1
Case 1 – Sue's Dialogic Interaction with ChatGPT in the First Session

Figure 2.2
Case 1 – Sue's Dialogic Interaction with ChatGPT in the Last Session

Figure 2.3
Case 2 – Ivy's Sample Dialogic Interaction with ChatGPT

Figure **3.1**
Participant Reported Cognitive Load by Prompt Type

Figure **3.2**
Relationship between ChatGPT Perceived Impact and Metalinguistic Awareness

Figure 4.1
Distribution of Response Types across the Top Three Categories of Feedback Received from ChatGPT

Figure 4.2
Distribution of Response Types across Participants

Figure 5.1
Writing Process and Feedback Loop

Figure 5.2
Comparison of Student-Generated and ChatGPT-Generated Linguistic Features Normalized Means

Figure 5.3
Comparison of Student-Generated and ChatGPT-Generated Lexical Diversity Measures

Figure 6.1
Video Conference between Classes

Figure 6.2
Example of Interaction with the Personalized ChatGPT

Figure 6.3
Example Showing Students' Adjustments toward More Specific and Concise Prompts

Figure 6.4
Screen Showing Students' Interaction with Google

Figure 6.5
Example Showing Enhanced Precision and Success in Student–ChatGPT Exchanges

Figure 6.6
Example of Students' Uptake of ChatGPT's Correction and Demonstrated Pragmatics Gains

Figure 8.1
Three Types of Human-AI Interaction (adapted from Hollnagel in Rasmussen & Andersen, 1991)

Figure 8.1
A Snapshot of a Humanoid Hyper-Realistic Avatar Video (from https://app.heygen.com/signup)

Figure 8.2
Descriptive Feedback Given Based on the Learners' Speaking Performance

Figure 8.4
Example Question-Level Feedback Provided to A Learner

Figure 9.1
Rubric for Assessing Content in CLIL Classroom

Figure 9.2
Detailed Prompt Structure of Training for Scoring Task 1 for ChatGPT

Figure 9.3
Variable Map for Dataset 1 (Humans + ChatGPT 1)

Figure 9.4
Category Statistics of the Rubric (Dataset 1: Humans + ChatGPT 1)

Figure 9.5
Probability Curb of the Rubric (Dataset 1: Levels 0–4)

Figure 9.6
Rater Measurement Reports of Dataset 1 (Humans + ChatGPT 1)

Figure 9.7
Example Response for Task 2 (Modified Based on the Student's Original Script) on the Left and the Researchers' Interpretation of the ChatGPT's Evaluation of the Response on the Right

Figure 10.1
Survey 1, Question 1: Number of SMEs Selecting 20 AI-generated Prompts as Suitable for Task 1 (Red) and Task 2 (Green) in the CILS B2 written production component (N=7)

Figure 10.2
Survey 1, Question 2: Number of SMEs Identifying 20 Writing Test Prompts as AI-generated (N=7)

Figure 10.3
Survey 2, Question 1: Histograms Showing Test Takers' Judgments of the Difficulty of Four Prompts (Prompt 1 and 2 Written by Humans; Prompts 3 and 4 AI-Generated) (N=63)

Figure 10.4
Survey 2, Question 2: Test-takers' Knowledge of AI (N=63)

Figure 10.5
Survey 2, Question 3: Test-takers' Usage of AI (N=63)

Figure 10.6
Survey 2, Question 4: Test-takers' Perception of AI Involvement in Writing Test Prompts (N=63)

Figure 10.7
Survey 2, Question 5: Sentiment Analysis of Test-takers Responses (N=39)

Figure 11.1
Scatterplot of Item Difficulty Estimates Between CTT Empirical Results and AI Models

Figure 11.2
Scatterplot of Item Discrimination Estimates Between CTT Empirical Results and AI Models

Figure 13.1
Mixed-Methods Research Design

Figure 13.2
Example of How Appraisal Framework was Applied to the Abortion Tweets

Figure 13.3
Application of Du Bois' Stance-Taking Framework to Analyze Twitter Posts on Abortion.

Figure 13.4
Example of How Stance Framework was Applied to the Abortion Tweets.

Figure 13.5
Frequency of Emotion Categories by Coder

Figure 13.6
Example of Full Agreement on Emotion Classification across Coders

Figure 13.7
Example of Disagreement on Emotion Classification across Coders

Figure 13.8
Frequency of Stance Categories by Coder

Figure 13.9
Example of Stance Disagreement between ChatGPT and Human Coders

Figure 13.10
Example of Disagreement in Emotion Classification between Human Coders and ChatGPT

Figure 13.11
Example of Disagreement in Stance Classification between Human Coders and ChatGPT

Editorial Team

Editors

Carol A. Chapelle
Iowa State University

Gulbahar H. Beckett
Iowa State University

Bethany E. Gray
Iowa State University

Assistant Editor

Duong Nguyen
Iowa State University

Editorial Review Panel

Sara Afifi
Iowa State University

Jeanne Beck
Iowa State University

Nergis Danis
Iowa State University

Sebnem Kurt
Iowa State University

Young-Joo Lee
Iowa State University

Febriana Lestari
Iowa State University

Cindy Lim
Iowa State University

Chris Litten
Iowa State University

Duong Nguyen
Iowa State University

Mostafa Ranjibar
Iowa State University

Andrias Susanto
Iowa State University

Contributors

Roula Aldib, *Georgia State University*

Danilo Calle, *Iowa State University*

Jean Young Chun, *Defense Language Institute Foreign Language Center*

Melinda Dooly, *Universitat Autònoma de Barcelona*

Mahdi Duris, *Iowa State University*

Onyemowo Angela Ejeme, *Oakbridge Foundation*

Robet Godwin-Jones, *Virginia Commonwealth University*

Volker Hegelheimer, *Iowa State University*

Sofie Johansson, *Göteborg University*

Rie Koizumi, *University of Tsukuba*

Sebnem Kurt, *Iowa State University*

Yusuke Kubo, *University of Tsukuba/Fukuoka University*

Lina Larsson, *Göteborg University*

Hongli Li, *Georgia State University*

Sabrina Machetti, *University for Foreigners of Siena*

Chad Marchong, *Georgia State University*

Paola Masillo, *University for Foreigners of Siena*

Maria Mont, *Universitat Autònoma de Barcelona*

Inyoung Na, *Iowa State University*

Tuba Özturan, *Erzincan Binali Yıldırım University*

Cristina Pardo-Ballester, *Iowa State University*

Giulia Peri, *University for Foreigners of Siena*

Prithvi N. Shrestha, *School of Languages and Applied Linguistics, Open University*

Andrias Susanto, *Iowa State University*

Saki Suemori, *University of Tsukuba*

Acknowledgements

The chapters in this volume were presented at the 2024 Technology and Second Language Learning conference at Iowa State University. The editors are grateful to the chapter authors who chose to share their groundbreaking research in this volume and then worked through the process of review and revision to make it a success. We also thank the many contributors who made the hybrid conference a success.

The organizing committee worked diligently throughout the process of developing the call for papers, fielding questions, developing expertise in implementing the technologies for hybrid delivery, managing the abstract submission and adjudication process, scheduling and logistics, as well as hospitality. We want to extend our sincere thanks to Mahdi Duris who masterfully managed the organization and communication for the conference. We also appreciated Mahdi's skills as he transitioned the project from conference to book by continuing to apply his expertise to organizing and communicating with prospective authors while onboarding the assistant editor and reviewers. We thank Cindy Lim for her expert management of the poster session and Danilo Calle for his efficiency and attentiveness in making sure that in-person participants had appropriate refreshments. We also thank members of the organizing committee for the many essential functions they carried out. The organizing committee consisted of the following volunteers:

Sara Afifi
Jeanne Beck
Gulbahar Beckett
Danilo Calle
Carol Chapelle
Mahdi Duris
Ali Ebrahimpourlighvani
Bethany Gray
Gi Jung Kim
Hwee Jean (Cindy) Lim
Duong Nguyen
Jim Ranali
Andrias Susanto

We also extend our thanks to the abstract reviewers who participated in the adjudication process resulting in the selection of papers and posters presented at the conference. The following people served as abstract reviewers:

Sara Afifi
Stacy Amling
Shireen Baghestani
Wren Bouwman
Junghee Byun
Danilo Calle
Kate Challis
Sarah Davis
Khoa Do
Ali Ebrahempour
Bethany Gray
Nazlınur Göktürk
Şebnem Kurt
Ewa Kusz
Erika Latham
Febriana Lestari
Hwee Jean (Cindy) Lim
John Levis
Sara Lastine
Amy Walton
MacKenzie Novotny
Duong Nguyen
Qi Qi
Quy Pham
Kristin Terrill
Erik Voss
Beata Walesiak
Amy Walton
Junghun Yang
Monique Yoder
Zoë Zawadzki

Agata Guskaroska
Angela Hakim
Droste Hennings
Trang Ho
Sarah Huffman
Gi Jung Kim
Luke Plonsky
Hannah Reischauer
Alif Silpachai
Andrias Susanto
Sinem Sonsaat-Hegelheimer
Burak Şenel

The conference committee was assisted by many volunteers who stepped forward as the conference approached to carry out important tasks such as chairing sessions. We want to thank these volunteers who helped in making the conference run so smoothly:

Sara Afifi
Ebtehal Asiri
Fatimah Aseeri
Mahdi Duris
Ali Ebrahimpourlighvani
Gi Jung Kim
Erika Latham
Hwee Jean (Cindy) Lim
Chris Litten
Inyoung Na
Duong Nguyen
Hannah Reischauer
Burak Senel
Andrias Susanto
Shuhui Yin

Finally, we gratefully acknowledge the institutional support that this project has received from Iowa State University through the Department of English under the leadership of Professor Volker Hegelheimer, from the College of Liberal Arts and Sciences, and from the Iowa State University Digital Press.

Chapter 1

Introduction to Researching Generative AI in Applied Linguistics

Carol A. Chapelle, Gulbahar H. Beckett, and Bethany E. Gray
Iowa State University

Almost two years after the public debut of ChatGPT, a group of applied linguists from around the world met virtually to share their research investigating a range of questions about generative AI (GenAI). Applied linguists study real-world problems in which language plays a role, including but not limited to issues in language education. The release of ChatGPT in 2022 expanded the set of language-related issues for applied linguists who were already investigating the many uses of language technologies in language learning, assessment, and research. Large language models, the backbone of GenAI, power qualitatively different tools from the digital tools of the past because of their capability to recognize and generate grammatical human-like language in response to human language input. The interfaces for tools such as ChatGPT provide access to anyone wishing to use GenAI. GenAI tools are available to users through a web browser, making them accessible to language students, teachers, researchers, and developers of materials for assessment and learning. The ubiquitous access to the power of GenAI creates the need to learn more about the actual uses of GenAI in applied linguistics and the quality of its language generation capabilities for these uses.

This volume begins to address this need with selected papers on the theme "Researching the Language and Use of Generative AI in Applied Linguistics" presented at the 2024 Technology for Second Language Learning hybrid conference held at Iowa State University, October 24 through 26. The chapters are organized into three sections, each on a significant theme in applied linguistics. The first two sections include chapters reporting on studies of GenAI use in language learning and in language assessment, respectively. The chapters in the third section examine the language capabilities of GenAI relevant to many aspects of applied

Cite as: Chapelle, C. A., Beckett, G. H., & Gray, B. E. (2025). Introduction to researching generative AI in applied linguistics. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 1-10). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.01
©2025 Chapelle, Beckett, and Gray. Published under a CC BY license.

linguistics research and practice, chief among them the need to adopt a critical eye toward GenAI performance.

Researching Language Learners' GenAI Use

The five chapters in the first section investigate how teachers and learners put GenAI tools to work in language learning. The contexts of language learning span locations, learners' characteristics, instructional conditions, and languages. What the research in this section shares is the interest in discovering how learners choose to use GenAI tools that the teachers and researchers make available to them in connection with their language classes. The first four chapters in this section examine how learners in different situations use ChatGPT in their writing processes as a tool for supporting revision of their writing and for formulating feedback to their partner in a peer evaluation. The final two chapters in this section each describe how GenAI was used in helping learners work with content in their respective target languages.

In Chapter 2, Tuba Ozturan and Prithvi Shrestha investigate the dialogic interaction with ChatGPT-3.5 of two adult learners of English in Türkiye wanting to improve their written English to meet their needs for advanced study and work. One had experience working with GenAI tools, but the other was a novice. The researchers observed the learners' interactions with ChatGPT-3.5 as they composed and revised a series of four argumentative essays. The learners began the revision process of the first essay by using a prompt formulated and tested by the researchers to instruct ChatGPT to respond like a TESOL teacher. Specifically, it was instructed to provide support through interactions including adaptive feedback on both local and global errors on a continuum from implicit to explicit while taking into account the need for developing metalinguistic knowledge and giving emotional support. After beginning the revision process using a prompt prepared by the researchers, the learners engaged in dialogic interaction with ChatGPT throughout the rest of the revision process. After each of the revision sessions, the learners shared the chat-log with the researchers, who conducted sessions eliciting the learners' reflections on their interactions with ChatGPT. The reflections revealed their motivations for seeking help beyond the feedback provided by ChatGPT. The interactions and revisions were interpreted as evidence about ChatGPT's capability for scaffolding users within their individual zones of proximal development (ZPDs) as defined in sociocultural theory and put into practice in dynamic assessment (Poehner, 2008). The study provides examples of the positive interactions that highly motivated individuals can engage in with a GenAI tool, the preparation in prompt development required to launch such interactions, and the methods teachers and researchers can use to better understand how their students use GenAI tools in their writing process.

In Chapter 3, Sofie Johansson and Lina Larsson also describe an investigation of the use of GenAI tools by professionals, in this case, instructors in higher education in Sweden who were motivated to increase their workplace language performance. The participants in the study were foreign-trained educators with a CEFR C1 level of Swedish who used Generative AI tools, including ChatGPT and Copilot, to support their learning of Swedish as a foreign language. The 16 learners participating in the research were teachers studying Swedish in a language course, where they were introduced to ChatGPT through an orientation session with instruction about its functionality, how to request targeted grammatical feedback, and practice in designing prompts.

Writing tasks were designed to engage learners in topics relevant to their uses of Swedish, including recounts of teaching experiences, reflection on educational challenges, and formulation of arguments to support a teaching method. Interested in developing a multifaceted understanding of the learners' use of GenAI and its role in their learning, the researchers grounded the study in three theoretical frameworks: Interaction Theory (Gass & Mackey, 2014), the Technology Acceptance Model (Liu & Ma, 2023), and Cognitive Load Theory (Sweller et al., 2011). Accordingly, data were collected using questionnaires, submitted assignments, and Chat-logs showing their interaction. These data were analyzed to find how learners' success in GenAI use for error correction in writing was associated with a combination of learner and situational factors. The results suggest that integrating ChatGPT into language learning requires a dual approach: instructional scaffolding to support learners in formulating effective prompts and adapting the technology to guide user interactions. The researchers concluded that explicit training in metalinguistic awareness and task-specific prompting can empower learners to engage with ChatGPT more effectively, ensuring the feedback they receive is both meaningful and actionable.

Chapter 4 reports an investigation of ChatGPT use by eight nonnative speakers of English enrolled in an ESL writing course at a university in the United States. The students had met the English proficiency requirement for university admission but then were placed into the ESL writing courses to improve their academic English. In this context, where many students take ESL writing courses, the availability of GenAI tools are expected to have an impact on learners' writing, the researchers, Inyoung Na, Mahdi Duris, and Volker Hegelheimer, wanted to learn how these students use ChatGPT to support or hinder their writing processes. After introducing the participants to the use of ChatGPT in the revision process, the researchers observed the learners during four revision sessions within a 3-week span. Their use of ChatGPT was documented in ChatGPT-learner conversation logs and screen recordings, and the researcher conducted semi-structured interviews to obtain data revealing learners' perspectives on their use of ChatGPT. The data showed variation in the learners' use of and perspectives on ChatGPT for revision, which were analyzed for their potential to contribute to learning. Positive interactions for learning were identified as instances where learners revised their writing in response to feedback from ChatGPT. This modified output, as defined in interactionist second language acquisition theory (e.g., Gass & Mackey, 2014), appeared in the revisions learners had made to content, organization, and sentence-level lexicogrammatical features. The variation in ChatGPT use was examined more holistically by examining case narratives of three participants who displayed distinct learner behaviors in their interaction with ChatGPT. One learner, who critically evaluated and selectively integrated ChatGPT's feedback to produce abundant modified output, showed the most beneficial patterns for autonomy and meaningful L2 development. In contrast, another uncritically adopted ChatGPT-generated language by copying and pasting it into their text without carefully modifying their own output. This study contributes to the emerging understanding of AI-assisted writing by identifying specific interaction patterns that may enhance or inhibit language development, providing insights for tailoring instructional strategies to maximize the potential of GenAI tools in L2 writing pedagogy.

Chapter 5 reports an investigation of the effects of ChatGPT on the peer review process in the same setting of ESL academic writing courses in higher education in the United States. The teacher-researcher Danilo Calle also examined learners' ChatGPT use, but prior to the observations, he had defined a specific role for ChatGPT to play in the peer review process: The reviewer was to consult ChatGPT prior to evaluating their peers' paper. Calle hypothesized that the use of ChatGPT to assist with the review would influence both the language and the scope of the reviewers' evaluation of their peers' papers due to the input the reviewer would receive from ChatGPT prior to producing their own feedback. In a qualitative study, he gathered the feedback generated by ChatGPT and the feedback produced by eight students who participated in three writing tasks after receiving ChatGPT-generated feedback on each one. The two sources of peer feedback (ChatGPT and peer reviewers) were each compiled into a corpus for analysis of linguistic complexity, lexical diversity, and evaluation focus. The results revealed two interesting findings. First, ChatGPT's feedback exhibited a higher degree of sentence complexity and lexical diversity while student feedback contained more epistemic markers and lexical repetition. Second, the comparison of the peer reviewers' language over the course of the three reviews suggested that their interaction with ChatGPT may have influenced the students' language, leading to increased nominalizations and a more informational style over time. This study demonstrates a productive use of ChatGPT in support of typical classroom writing activities that may contribute to the learners' engagement in the peer review process and their language for doing so.

In Chapter 6, Maria Mont, Melinda Dooly, and Onyemowo Angela Ejeme described their exploration of the role GenAI played in a telecollaborative learning project for English language learners in Spain and Nigeria. "The Young Activists Project" carried out over three months connected 52 learners in Catalunya, Spain (ages 9-10 years-old) with seven in Abuja, Nigeria (ages 10-14). This initiative combined AI with geographically distributed, project-based learning with the aims of enhancing English language skills, digital literacy, and global awareness among young learners. A customized GenAI platform was developed for English language support throughout the multiple tasks in the project. Through multimodal interaction analysis applied to video recordings and screen recordings, the study examined how AI-powered tasks promoted language use and digital awareness during the project. Findings indicated that the young learners developed critical digital literacy skills by learning to refine their inputs for effective AI interactions, and they engaged with English more actively than they typically did in traditional classroom settings.

In Chapter 7, Cristina Pardo-Ballester explores how novice translators in a college-level Spanish – English translation course used GenAI tools to facilitate their interpretation of the humor in jokes requiring an understanding of cultural context and wordplay. Seventeen student translators with an intermediate-mid Spanish proficiency level carried out two humor translation tasks, in which they reviewed a joke (consisting of text in Spanish and an accompanying picture) alongside literal translations performed by traditional machine learning tools. Students then used one or more GenAI tools (ChatGPT, Gemini, and/or Copilot) to assist in crafting their own English translations of the jokes. Relying on Attardo's (2020) General Theory of Verbal Humor, Pardo-Ballester evaluated the final translations and examined them alongside student reflections

and prompt-output logs documenting how participants interacted with GenAI tools. Findings indicated the importance of prompt engineering – translators who were able to craft appropriate prompts to engage with GenAI tools to comprehend the deeper meaning of the jokes, place them within specific cultural contexts, and understand wordplay leading to humor were ultimately able to produce more successful translations. This study reinforces the need to train users on effective prompting techniques to address specific language-related tasks, such as humor translation.

Researching GenAI in Language Assessment

GenAI raises many new possibilities for increasing efficiency in language assessment, and each one requires extensive research to assess the value attained through the efficiency relative to any potential negative consequences. The four chapters in this section investigate four uses of GenAI in language assessment ranging from evaluating test takers' performance to developing testing materials and validation studies conducted through simulation of test takers' performance.

In Chapter 8, Andrias Susanto reports the results of a study investigating the use of a GenAI tool for assessing learners' speaking ability in a privately-run IELTS (International English Language Testing System) preparation program taught by TESOL professionals who serve as tutors to help students prepare for the speaking section of the IELTS. As part of the preparation, the tutors administer practice speaking tests modeled on what the students will be asked to do on the IELTS. After experiencing the practice IELTS administered by the tutors, 11 students accepted the invitation to participate in the research by taking another practice IELTS speaking test, this time administered by the GenAI tool with an interactive speaking avatar configured to play the role of the human examiner. The goal of the research was to discover the learners' perceptions of their experience being tested by the AI agent that interacted with them in conversation, scored their performance according to the IELTS rubric, and gave them more specific feedback. Overall, the participants praised the efficiency, objectivity, and detailed feedback provided by the AI agent. Based on the agent's pause detection and word count analysis, it provided different types of feedback from that of the human tutors. The learners also recognized the personalized, adaptive, and in-depth feedback that the human tutors provided them because of the way they created a supportive learning environment. The participants' comments suggested promise for a hybrid test preparation program that offers learners the benefits of practice speaking assessments administered by the human tutor and by the AI agents at various points during the program.

In Chapter 9, Rie Koizumi, Saki Suemori, and Yusuke Kubo extended the study of GenAI use for assessment by investigating ChatGPT-4's capabilities in ratings the content of test takers' English performance. Students were in a content and language integrated learning (CLIL) course where their final presentations were scored on an analytic rubric requiring evaluation of both the content and the language. For this exploratory study, the researchers obtained 52 presentation scripts from 37 university students who had taken the course. They prompted ChatGPT-4 using an analytic rubric with content as one of the criteria, an explanation of the rubric, example scripts, and benchmarking scores to request the ratings for each presentation script. With the understanding that ChatGPT cannot be assumed to display consistent

performance, this procedure was repeated three times to obtain three ratings from ChatGPT-4, and many-facet Rasch was used to compare these ChatGPT's ratings with those given by seven humans on the same rubric. The results showed that seven human raters and ChatGPT-4 produced consistent scores, with ChatGPT-4 tending to produce slightly lenient ratings and fewer biases suggesting promise for the use of ChatGPT for rating content in language assessments.

In Chapter 10, the researchers, Giulia Peri, Sabrina Machetti, and Paola Masillo, describe their investigation of GenAI for producing test content for the Certification of Italian as a Foreign Language (CILS) exam at the B2 level. They carefully designed prompts to instruct ChatGPT-4 on the necessary characteristics of the test content based on models of materials created by humans, and then they instructed it to develop ten writing task prompts for each of two writing tasks for the test. The twenty writing task prompts were evaluated by expert test developers using a survey constructed to elicit their judgments of the appropriateness, representation of the B2 construct, and AI detectability. The same twenty writing task prompts were judged by test takers who responded to a survey after they had taken the writing test containing some of the prompts. Neither group of evaluators was aware that all of the writing task prompts had been created by GenAI. Based on the evaluations, the researchers saw promise for AI-generated writing task prompts, particularly noting their clarity, thematic relevance, and perceived difficulty. However, they also found that some would need to be refined to better display an appropriate level of language complexity and adhere to the CILS format.

Chapter 11 also describes an investigation of ChatGPT's use in procedures carried out in the development phases of a language assessment. In two exploratory studies, the researchers, Hongli Li, Roula Aldib, and Chad Marchong, tested ChatGPT's use for shedding light on issues in test development and validation of a reading test. In their first study, ChatGPT was used to estimate item difficulty and discrimination for items on a reading comprehension test based on its evaluation of the item content. Item difficulty and discrimination statistics are used in test development to assemble multiple test forms that are likely to be equivalent in difficulty and in their capability to discriminate among examinees. These statistics are normally obtained by piloting a draft version of the test on a sample of examinees and calculating the numbers of students responding correctly to each item (difficulty) and the relationship between correct responses and overall scores for each item (discrimination). If an AI tool were able to estimate these statistics based on its language analysis, without actually testing a sample of examinees, test development could be considerably streamlined. The second study of a reading test used ChatGPT to investigate the extent to which reading items could be answered correctly without access to the reading passage. This analysis was designed to address a fundamental question about reading items: Do examinees' correct responses result from comprehension of the text in the test and not from the examinees' background knowledge? ChatGPT's performance based on its nearly infinite knowledge would seem to provide a good test of the role of background knowledge in correctly responding to questions. In both studies, the results provided by ChatGPT were compared to those obtained from human test takers to evaluate the utility of ChatGPT's performance in each study. The authors interpret the findings as suggesting some potential for ChatGPT as a tool for facilitating reading test development and validation at the

early stages to engage test developers in constructive reflection and dialog about draft test items. However, the results also demonstrate the limitations of ChatGPT's performance in these roles.

The four assessment studies investigating prospective roles for GenAI tools in language assessment begin to reveal the variety of uses assessment specialists see for GenAI as well as the methods required for evaluating how well the GenAI performs. In each case, the assessment context and use were precisely described. Based on the prospective role of GenAI for the specified purpose, the researchers identified the most informative path for evaluation, which in each case included human judgments systematically gathered from knowledgeable, relevant participants in the assessment process. The statistical and conceptual tools of assessment specialists were applied to obtain precise data needed to address the evaluation. Because of the extensive use of assessments in applied linguistics research, some of the findings are more broadly relevant for language learning and research. The final three chapters, however, focus the research directly on the language capabilities of GenAI.

Researching GenAI's Capabilities

The final section of the volume mounts challenges for GenAI in the face of the human dimensions of language that are particularly important in applied linguistics, including regional language variation, expressions of subjectivity, and pragmatics. These three aspects of language use at the core of applied linguistics work extend beyond the regular structural linguistic patterns of language use that GenAI demonstrates so impressively. Regional language variation manifests the language conventions that are not regular across all speakers of a language, but that mark language users' identity as a member of a particular regional speech community. Expressions of subjectivity convey human thoughts and emotions that are understood in human-human communication when a level of intersubjectivity is achieved by interlocutors communicating within a certain time and place. Expressions of subjectivity therefore have a dimension of pragmatic meaning, which GenAI appears to exhibit in some instances of its language use. However, the range of pragmatic meanings in human language extends beyond the contextually tuned choices of language users to include the choices emanating from culturally conditioned conventions that define what a language user can mean within their context of language use.

In Chapter 12, Jean Young Chun tests ChatGPT's adeptness with language variation in a study investigating its capability for recognizing and generating North vs. South Korean dialects. ChatGPT-4o was presented with recognition tasks including dialect distinction and identification of North Korean words and their meanings in 200 passages, half representing each of the two dialects. It was also prompted to perform production tasks, such as translating the South Korean dialect into the North Korean dialect. The study found that ChatGPT-4o was highly capable of recognizing each of the dialects and identifying the meanings of North Korean words. In contrast, its performance in translating the South Korean dialect to North Korean was limited. The study demonstrates the need to investigate rather than assume that GenAI tools such as ChatGPT have the capacity to handle the language variation that exists in the real world.

In Chapter 13, Şebnem Kurt and Mahdi Duris describe their study that seeks to discover how well ChatGPT-4o performs in discourse analysis research for coding the subjective language expressing emotion and stance. The study compared two human analysts' and ChatGPT's coding of a dataset of 100 tweets posted on the topic of abortion from Boston, Massachusetts, collected between September 2022 and January 2023. Consistent with common practices in qualitative discourse analysis, emotion was defined based on Martin and White's (2005) appraisal framework, focusing on AFFECT to code sentiments expressing positive, negative, and neutral emotions. Stance was defined using Du Bois' (2007) stance-taking analysis to code evaluation and alignments, indicating pro-abortion, anti-abortion, or neutral stances. Despite the knowledge supplied to ChatGPT-4o about the coding schemes, it did not achieve an acceptable level of agreement with either of the two human coders for use in most applied linguistics research. Nor was the overall reliability across the two human coders and ChatGPT acceptable for coding emotion or stance. The limited agreement between ChatGPT and raters and the overall inadequate reliability are interpreted in view of the fact that the two human coders also achieved only modest levels of agreement as well (Kappa=0.355 for emotion and Kappa=0.402 for stance). The limitations in agreement achieved by the human coders prompted the researchers to examine ChatGPT's performance on the two subsets of the tweets that the two humans had agreed upon. In these data, even though the researchers were able to identify subjectivity in the language that was apparent to both of them, it was not necessarily apparent to ChatGPT. The study underscores the implication stated in Chapter 12 about the need for careful evaluation of any coding performed by GenAI tools in applied linguistics research.

A most fitting closing chapter to *Researching GenAI in Applied Linguistics*, Robert Godwin-Jones tackles the challenge facing applied linguists going forward. Based on superficial observations that GenAI tools such as ChatGPT perform linguistically in such a human-like fashion, too many people conclude that AI can replace human interlocutors in language learning activities. The seemingly apparent readiness of GenAI tools to stand in for human performance challenges applied linguists to put forward a sophisticated analysis of the multi-layered systems of language, knowledge, and experience comprising human interactional competence. Such an analysis is needed for understanding the meaning of critical AI literacy for language teachers and learners. AI literacy in applied linguistics needs to encompass an appreciation for both the remarkable feat of building the large language models underlying the conversational performance of GenAI and its limits. Godwin-Jones' analysis of the adequacy of existing approaches to pragmatic theory and the research testing the pragmatic competence of GenAI tools provides a good foundation for continued research working toward critical AI literacy.

Does GenAI Work in Applied Linguistics?

The chapters in this volume show a sample of the uses that applied linguists have in mind for GenAI and some of the ways that they are investigating how well it works. Unlike the popular media speculation on the general utility of the language generation capabilities of GenAI tools, applied linguists require evidence that the tools support their intended purposes with a level of technical accuracy appropriate to their task goals. The types of tasks described in parts one and two of the volume are categorized as language learning and assessment, respectively, but within

these overall groupings, each project had a specific use described for the GenAI and a means of describing and evaluating the performance of the GenAI tool. In the four studies of learners' GenAI for writing support (Chapters 2-7), the researchers examined data showing the interactions between each of the learners and the GenAI tool. In each case, the interactions were interpreted relative to criteria derived from previous theory and research in applied linguistics to provide a frame of reference for evaluating the success of the learners in interaction with the tools. The studies investigating the use of AI tools in language pedagogy described specific uses for the AI that provided a basis for their evaluation. In all these cases, the question of whether the GenAI worked was not about the tool alone. It had to address the way the learners and teachers used the tools.

Similarly, research on the use of GenAI in language assessment required specification of the intended use of the tools in addition to data showing their performance for their specified purposes. The data in each of the four chapters in this section (Chapters 8-11) is unique to the testing purpose and the critical questions for the success of the use of GenAI. When the GenAI was to be used in practice speaking assessments (Chapter 8), a critical question was how the learners would perceive the GenAI agent as an examiner. When GenAI was considered as a rater of test takers' performance, the researchers wanted to know how its consistency, severity, and degree of bias would compare to those characteristics of ratings produced by human raters. When GenAI was used to construct writing prompts for a test, the research required human test developers and prospective test takers to judge the quality of the generated prompts. When GenAI was used to estimate item statistics and to respond to test items, its performance was judged against results obtained from human test takers. In each case, the critical areas for examination were carefully crafted to yield appropriate evaluations of performance, and in each case, the evaluation includes human performance, judgements, or both.

The chapters in the final section tackle the hard problems for GenAI—characteristics of human language use in the real world that include regional variation, expression of subjectivity, and pragmatics. In these final chapters, the authors describe the role of human judgment and analysis in evaluating GenAI's capabilities. In each of the chapters, the authors reveal some of the successes of GenAI, but also recognize the limitations. In a world that is focused on the successes of GenAI—of which there are many—the chapters in this volume present a more measured view by differentiating among GenAI uses in applied linguistics and engaging their respective methods of evaluation.

References

Attardo, S. (2020). *The linguistics of humor: An introduction*. Oxford University Press. https://doi.org/10.1093/oso/9780198791270.001.0001

Du Bois, J. W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (pp. 139-182). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.07du

Gass, S. M., & Mackey, A. (2014). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition* (pp. 180–206). Routledge. https://doi.org/10.4324/9780203628942

Liu, G., & Ma, C. (2023). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching, 18*(2), 125–138. https://doi.org/10.1080/17501229.2023.2240316

Poehner, M. E. (2008). *Dynamic assessment.* Springer. https://doi.org/10.1007/978-0-387-75775-9

Martin, J. R, & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan. https://doi.org/10.1057/9780230511910

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

Chapter 2

ChatGPT's Potential to Enhance Language Learners' Proximal Development through Dialogic Interaction

Tuba Özturan[1] and Prithvi N. Shrestha[2]
[1]*Department of Translation and Interpretation, Erzincan Binali Yıldırım University*
[2]*School of Languages and Applied Linguistics, The Open University*

Dynamic Assessment (DA) is predicated on the dialogic interaction between teachers and learners, emphasizing the fusion of assessment and instruction and revealing learners' Zone of Proximal Development. Despite its promising contributions, DA in actual classrooms can be overwhelming for most teachers. To address this issue, this original study showcases the interaction between learners and ChatGPT-3.5. Participants were L2 writers in an EFL setting, and the targeted genre was argumentative writing. For four weeks, the learners produced argumentative texts on different topics, and they used ChatGPT as a personalized feedback and instruction provider. They initially used the same prompt, and following the ChatGPT-generated feedback, participants were free to seek further assistance. After each session, they shared the chat-log with the researchers, and reflection sessions were conducted to delve deeper into the learners' interaction and their underlying reasons for seeking additional help. The findings reported ChatGPT's potential to foster users' ZPD and enhance their improvement in writing since each participant's interactional patterns, microgenesis, and transfer abilities were in accord.

Cite as: Özturan, T., & Shrestha, P. N. (2025). ChatGPT's potential to enhance language learners' proximal development through dialogic interaction. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 11-39). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.02
©2025 Özturan and Shrestha. Published under a CC BY license.

Introduction

Dynamic Assessment (DA) is a learning-oriented assessment approach based on the dialogic interaction between teachers and learners, emphasizing the integration of assessment and instruction through adaptive feedback. This process-based method aims to uncover learners' Zone of Proximal Development (ZPD) by establishing a personalized learning-assessment setting (Poehner, 2008). Its theoretical background relies on Vygotsky's Sociocultural Theory and Feuerstein's Mediated Learning Experience (Feuerstein et al., 2010; Vygotsky, 1980). Accordingly, DA asserts that adopting a past-to-future assessment approach and assessing merely learners' current knowledge might not be sufficient to unpack their proximal development in the future. Rather, embracing a goal-oriented dialogic interaction that is based on an implicit-to-explicit feedback loop and including teachers and learners as active participants can yield each learner's present-to-future potential, namely ZPD (Lantolf & Poehner, 2004; Poehner, 2008). This goal-oriented interaction can rely on either interactionist or interventionist models. Poehner (2008) states that the former refers to providing feedback and mediation in an ad hoc manner while the latter refers to adopting pre-defined hierarchical mediational moves in an implicit-to-explicit continuum based on the scale reported by Aljaafreh and Lantolf (1994).

Being an iterative, systematic, process and interaction-oriented approach, DA encompasses two salient facets: instruction and assessment. On the one hand, teachers are expected to detect each learner's matured abilities (Zone of Actual Development) and needs followed by fine-tuning feedback that aligns with the learner's needs through an implicit-to-explicit feedback loop (Aljaafreh & Lantolf, 1994). Concurrently, learners are encouraged to articulate the rationale behind their responses, which is critical for examining the underlying reasons for erroneous output and their cognitive development in addition to providing subsequent feedback, if required (Poehner, 2008). On the other hand, tracking learners' microgenesis (tiny moves of improvement) and their ability to transfer knowledge over time (transcendence) are also important (Tzuriel, 2013). In this aspect, teachers are responsible for detecting any change in the nature of learners' errors as well as any change in feedback type both within a DA session and across DA sessions. This assessment methodology allows for predictions regarding learners' future performances (Shrestha, 2020).

DA was employed in different language learning settings, including influential studies in L2 writing. Accordingly, one-on-one DA was found to be an effective approach to teach Spanish interrogatives (Davin, 2016) and Wh- question formation in Spanish (Davin & Donato, 2013) to young learners and to have a long-term impact on EFL learners practicing English tense-aspect system (Infante & Poehner, 2019) and English derivational affixes (Poehner & Leontjev, 2020). Adopting online asynchronous interactionist DA in an academic writing setting, Shrestha (2020) illustrated participant learners' mediational moves, reciprocity acts, and near and far transfer abilities in terms of producing macro-themes and hyper-themes. Poehner and Yu (2021) also further investigated interactionist DA in an academic writing setting by embedding rubric use and yielded participants' differential needs in rhetorical and linguistic aspects. Özturan and Uysal Gürdal (2022) adopted interactionist DA to explore two multilingual students'

microgenesis and transfer abilities in L2 writing, both in organizational and linguistic aspects. Overall, these studies highlighted the salience of establishing a goal-oriented dialogic interaction with learners to create a personalized learning and assessment setting, among others.

DA presents promising practices and contributions to language education, particularly within L2 writing classes, such as increased motivation towards writing, teachers' supportive role, decreased anxiety, personalised learning and assessment setting, and tracking progress over time (Özturan et al., 2023). However, it also poses significant challenges for educators, especially in large classes and/or those with multilingual learners. In such contexts, it becomes difficult for educators to systematically monitor learners' progress over time and to provide appropriate forms of mediation systematically and iteratively (Davin & Herazo, 2020), which warrants blending computer technology with DA principles (Poehner & Van Compernolle, 2020).

Today, computers have tremendous improvements thanks to artificial intelligence, including Large Language Models. A Generative Pre-trained Transformer (GPT) is a type of Large Language Model (LLM) that is pre-trained on massive amounts of data through data mining and web crawling, to learn from different sources, such as Wikipedia, books, and articles and then fine-tuned to specific tasks. GPTs are based on transformer architecture and self-attention mechanisms, which enable them to interpret prompts to produce human-like texts, including summaries, responses to questions, and writing in different genres and languages, in addition to engaging in interaction with users (Barrot, 2023; Godwin-Jones, 2021; Kasneci et al., 2023). ChatGPT, the most widely known form of GPTs, is an AI-driven chatbot developed and released by OpenAI in November 2022.

Such remarkable advances in AI have garnered great interest from both researchers and educators who are interested in adopting technological tools in education and in investigating their impact on learning and assessment. Particularly, the capacity to produce human-like texts has been investigated through diverse paths in language education. One such path that is related to this study's scope is exploring ChatGPT-driven feedback and its impact on L2 writers' revision behavior (Koltovskaia et al., 2024; Ziqi et al., 2024). Another is investigating the comparable effect of ChatGPT and Pigai (an Automated Writing Evaluation tool) on L2 writers' writing skill and their ideal L2 writing self (Shi et al., 2025). However, dialogic interaction and adaptive feedback are under-researched in these studies.

As an emerging field, many studies have been carried out to unpack whether AI is leading and will lead to new teaching and assessment methods (Wang, 2024). However, to our knowledge, there is a paucity of related research focused on the fusion of GenAI and DA despite the growing interest in both in language education. DA's contributing promise to diagnose and enhance language learning through adaptive and personalised feedback over time has been illustrated in many studies, but they are mostly limited to grammar (e.g., Davin, 2016; Infante & Poehner, 2019). Therefore, on the one hand, DA's potential to improve writing quality both in global (organization, content, cohesion) and local (grammar, vocabulary choice, spelling, and punctuation) aspects (Lundstrom & Baker, 2009) needs to be investigated. On the other hand,

despite DA's contributions, it has some challenges for educators with its systematic, personalized, and iterative nature (Davin & Herazo, 2020). In this regard, GenAI's mechanism to have a dialogic interaction with end-users and to provide personalized adaptive feedback could help educators who aim to adopt DA in their classes. The lack of research in this area warrants further investigation to carefully examine the power of the fusion of GenAI and DA.

Grounded on the promise suggested by related studies, this chapter significantly contributes to the field by showcasing the interaction of Turkish learners of English with ChatGPT, an underexplored dimension in DA. The initial prompt was arranged to base the dialogic interaction on DA principles (i.e., implicit to explicit feedback as in Aljaafreh & Lantolf [1994]) and on the global and local errors in writing. The following research questions guided this chapter:

1. What are the interaction patterns of participants with ChatGPT-driven adaptive feedback pertaining to both global and local dimensions of writing?
2. To what extent do participants exhibit microgenesis within a task and subsequently transfer that knowledge across different tasks?
3. What reflections do participants provide regarding their experiences with ChatGPT-driven dialogic assessment sessions?

Methods

Participants

This original qualitative case study engaged two Turkish learners of English as participants (Sue and Ivy, pseudonyms) to illustrate their L2 writing and interaction with ChatGPT experiences holistically (Mackey & Gass, 2005). Sue, 36 years old, works as a civil servant and is preparing for the TOEFL examination to pursue an MBA in the UK. She reported prior experience with various artificial intelligence (AI) tools, including ChatGPT, AI Write, and DeepL. Conversely, Ivy, a 25-year-old lawyer, has no previous experience utilizing AI tools. Both participants are employed within the same governmental organization in Türkiye and share the objective of enhancing their written proficiency in English.

Data Collection

Case study research design, enabling data collection in an ecological and detailed setting through various information sources (Creswell & Poth, 2018), was adopted in this study. The data were gathered through the participants' chat logs and reflections to examine their interaction with ChatGPT. The purposed was to investigate the participants' proximal development over time through interaction patterns and error natures, and unveil their engagement through reflection notes in detail.

The data were collected in four sessions, focusing on argumentative writing as the targeted genre of this study, a genre we had previously researched in our DA research (Özturan

et al., 2023). In each session, a different argumentative writing topic was assigned to the participants, who composed their texts without the aid of machine translation or artificial intelligence tools. Upon completing their written texts, participants logged in to their ChatGPT (ChatGPT 3.5, free version) accounts (Sue already had an account, while Ivy created a new one). Subsequently, they utilized the prompt provided by the researchers and copied their written texts to initiate a dialogue with ChatGPT. The primary objective of the initial prompt was to facilitate dialogic interaction with each user and to emphasize error analysis based on global and local aspects separately. To ensure the efficacy of ChatGPT in delivering adaptive and scaffolded feedback while implementing principles of DA, the initial prompt was carefully tested and revised by the researchers prior to its assignment to the participants (Appendix B). By doing so, first a prompt was generated with all details that would guide ChatGPT to behave like a TESOL teacher, to implement DA and provide adaptive feedback in an implicit-to-explicit feedback continuum for both global and local aspects-based errors, to ask for the end-users their justifications for any error treatment or difficulty in it, and to provide emotional support for the users. Then, the first researcher wrote an argumentative paragraph, with some errors on purpose, and tested the prompt's efficiency. The first test session displayed that ChatGPT showed all errors at the same time, so the prompt was revised to make ChatGPT refrain from displaying all errors at the same time. Afterwards, the revised prompt was tested on a new chat log by the same researcher. Since ChatGPT succeeded in providing adaptive feedback for each error separately during interaction, the revised prompt was assigned to the participants.

For each session, the participants were required to fill in a form designed for DA-oriented interactions (as an example, see Appendix A). This was a personal form designed on Google Docs by the researchers and shared with each participant through their emails, and the form included the steps that the participants would follow for each session, the assigned topic, and the personalized prompt that each participant would use to start the dialogue with ChatGPT. Moreover, the participants were required to note how much time they spent writing the text, what kinds of problems they encountered while writing the text, what their reflections were, and their chat log links. After each session, the first researcher analyzed the chat logs, reflections, and any challenges experienced by the participants. Then, the subsequent personalized prompts were prepared on each participant's Google Docs form and sent to the participants via email.

In the first week, both participants utilized the same prompt. However, for the subsequent sessions, the prompts were tailored individually for each participant (Appendix B). This modification considered each participant's reflections, chat log history, and any challenges encountered to target their ZPDs. For instance, Sue employed the same prompt for both the first and second sessions. However, in the second session, the researcher noted that ChatGPT failed to provide scaffolded feedback; instead, it directly corrected errors and supplied explanations for the corrections when Sue struggled to fix the erroneous output. Moreover, Sue sought advice on how to improve her writing by posing questions to ChatGPT. Consequently, a personalized prompt was developed for Sue for the third session, which she also utilized in the final session.

Regarding Ivy, she expressed a preference for reviewing the entire revised version of her text at the conclusion of the session. Consequently, the prompt was adjusted accordingly for the

second session. In the third session, the researcher observed a comparable issue encountered by Sue; ChatGPT was unable to effectively advance the interaction and provide adaptive feedback, particularly concerning local errors. This motivated the researchers to fine-tune the prompt once more. Ivy utilized this modified prompt during the third and final sessions.

As the final step in data collection, both participants were requested to share their insights by writing on their Google Docs forms regarding the adoption of ChatGPT-driven dialogic interactions in relation to their errors. Although a comprehensive reflection was required from the participants following each session to delve into their session-specific experiences in detail, they were also asked to provide a general evaluation of their use of GenAI at the end of the study.

Data Analysis

Data analysis was threefold. First, the analysis examined the interactional patterns between the participants and ChatGPT-3.5 by identifying the type of dialogic feedback and learner response. This examination investigated and categorized ChatGPT's potential to provide adaptive feedback within an implicit-to-explicit feedback loop (Aljaafreh & Lantolf, 1994; Shrestha, 2020), as well as the participants' reciprocal responses to the feedback, focusing on both global and local errors. Furthermore, a comparison was made between the machine-driven feedback loop and the human-generated feedback loop by referencing prior notable research findings. Second, the study explored the participants' abilities to exhibit microgenesis within a task and to transfer knowledge to subsequent tasks. To achieve this, the nature of each participant's errors, analyzed in both global and local contexts, was assessed across four sessions, followed by presenting dialogic interaction cases to illustrate their microgenesis. Lastly, the participants' written reflections were documented and thematically analysed to further interpret and triangulate the data concerning their proximal development.

Results and Discussion

RQ1: What are the Interaction Patterns of Participants with ChatGPT-driven Adaptive Feedback Pertaining to both Global and Local Dimensions of Writing?

The first research question aimed to yield the interactional patterns of the two participants with ChatGPT, and we presented findings first, followed by their discussion in relation to the previous studies. However, before examining these patterns, it is essential to explore and discuss ChatGPT’s potential to enhance adaptive feedback within an implicit-to-explicit feedback loop, particularly concerning both the global and the local aspects of writing as revealed in this study. The interactional data analysis conducted across four sessions indicated that ChatGPT integrated various feedback approaches within single interactions and exhibited a limited number of turns—typically confined to three or four—to customize feedback according to the two participants' responses (Table 2.1). Regarding global aspect-based errors, the interaction generally started with ChatGPT identifying the location of the error, specifying its nature (such as clarity, quality of the topic sentence or thesis statement, and argumentation), and prompting the participants to

articulate the rationale behind any proposed error correction. In instances where the participants demonstrated a lack of autonomy in addressing errors and verbalizing their reasoning, ChatGPT sustained the interaction either by providing an example or a corrected/improved version of the erroneous segment, subsequently encouraging the participants to investigate the underlying reasons, or by offering a detailed explanation of the error. If the participants remained unable to exhibit full autonomy following any of these mediational interventions, ChatGPT delivered an explicit error correction alongside an explanatory note. Overall, ChatGPT demonstrated the capacity to offer adaptive feedback within an implicit-to-explicit feedback loop targeting global dimension-oriented errors, although the frequency of turn-taking was notably limited.

In contrast, ChatGPT's primary approach for addressing errors related to local dimensions was predominantly constrained to providing explicit feedback. This included initiating interactions by identifying the erroneous component through indicating the nature of the error, or by refining the focus on the erroneous element accompanied by an explanation of the correction. In instances where the participants did not address the error, ChatGPT intervened by directly correcting the mistake and providing a corresponding explanation.

Regarding the two participants' acts of reciprocity, they predominantly identified and addressed local aspect-related errors through offering explanatory notes. However, their responses to the feedback concerning global aspect-related errors exhibited greater variability. For instance, in certain instances, Ivy was able to provide a corrected version accompanied by an explanation. Conversely, Sue presented an improved version without offering an explanation, or she articulated the reasons for the error but did not correct it. Additionally, there were occasions in which both participants asked for recommendations (for sample interactions, please see Appendix C).

Overall, our findings are in accord with the concepts of human-generated feedback loops and learner reciprocity in the literature; however, some variations are evident. Table 2.2 presents a notable DA study regarding the human-generated feedback loop and learner responsiveness: Shrestha (2020) concentrated on evaluating students' academic writing skills through dynamic assessment. As indicated in the table, the human mediator provided more diversified feedback and mediational moves compared to machines, and learners displayed similar tendencies. One possible explanation for this discrepancy may relate to the advantages of in-person or human-to-human interaction, which facilitated multimodal engagement and tailored feedback moves that addressed learners' diverse needs and reciprocal interactions. Furthermore, ChatGPT integrated various feedback moves within a single response, which may pose challenges for some learners, particularly those who are not proficient in the target language. However, it is important to note that the effectiveness of ChatGPT is largely related to the quality of the prompts provided. Consequently, with careful attention to prompt formulation, ChatGPT has the potential to assist teachers in implementing dynamic assessment effectively.

Table 2.1

ChatGPT-Generated Feedback Loop and Users' Reciprocity Acts

Global Aspects		Local Aspects	
Feedback Loop by ChatGPT	Reciprocity Acts of Users	Feedback Loop by ChatGPT	Reciprocity Acts of Users
ChatGPT narrowed down the location of error, identified the error nature, and guided the user to treat the error and articulate the reasoning. ChatGPT displayed an improved version of the erroneous sentence. It also guided the user to verbalize the reasoning of this revision. ChatGPT highlighted the location of phrases/words that spoiled the clarity of topic sentence/thesis statement. OR ChatGPT guided the user to think about the location of the topic sentence/thesis statement. ChatGPT provided an explanation for the error correction and displayed the correct form.	User could provide the correct/improved version and verbalize the reasoning for it. User could provide a revision but was unable to articulate reasoning for it. OR User could explain the reasons for the error but was unable to treat the error. User asked for a recommendation. User explicitly mentioned the lack of knowledge and asked for an explanation.	ChatGPT pointed out the erroneous part and mentioned the error nature, such as grammar, spelling, or punctuation. ChatGPT directly corrected the error and explained the reason for it.	User could find error/s, fix it/them, and articulate the reasoning. User asked for help.

Table 2.2

Human-Generated Feedback Loop and Users' Reciprocity Acts (Shrestha, 2020, pp. 119 – 134)

Feedback Loop by Human Mediator	Reciprocity Acts of Writers
Clarifying the task	Unresponsive
Accepting a response	Imitating the mediator
Locating part of the text needing improvement	Responding incorrectly
Providing metalinguistic clues	Incorporating feedback
Exemplifying or illustrating	Overcoming problems
Providing a choice of possible solution(s)	Using mediator as a resource
Providing the correct solution	Rejecting the mediator's feedback
Explaining the problem	Asking for clarification
Asking learner to clarify meaning	Checking conceptual understanding with mediator
Showing affect	Asking for content clues
Asking learner to identify the problem	Identifying the problem
Providing content clues	Explaining the problem
Rejecting the response with explanation(s)	Evaluating mediator feedback
Asking to consider a possible solution	Self-assessing
Checking conceptual understanding	Suggesting a possible solution
	Verbalize conceptual understanding

RQ2: To What Extent do Participants Exhibit Microgenesis within a Task and Subsequently Transfer that Knowledge across Different Tasks?

Learners' ability to showcase tiny moves of improvement within a DA session (microgenesis) and to transfer that improvement to subsequent sessions and tasks (macrogenesis) are of importance to examine their ZPDs (Poehner, 2008). In this regard, we first analysed the error natures of both participants across four sessions and examined any decrease and/or change in them. Then, we focused on their dialogic interaction to reveal their microgenesis. Combining both analyses might illustrate the participants' ZPD.

Table 2.3 presents Sue's errors over four sessions. It is evident that Sue had consistent errors across the sessions, such as having an unclear argument in the text and choosing the inaccurate word. The retention in this consistency might show Sue's limited ZPD over these segments since she was in need of receiving assistance and feedback across all sessions.

Table 2.3

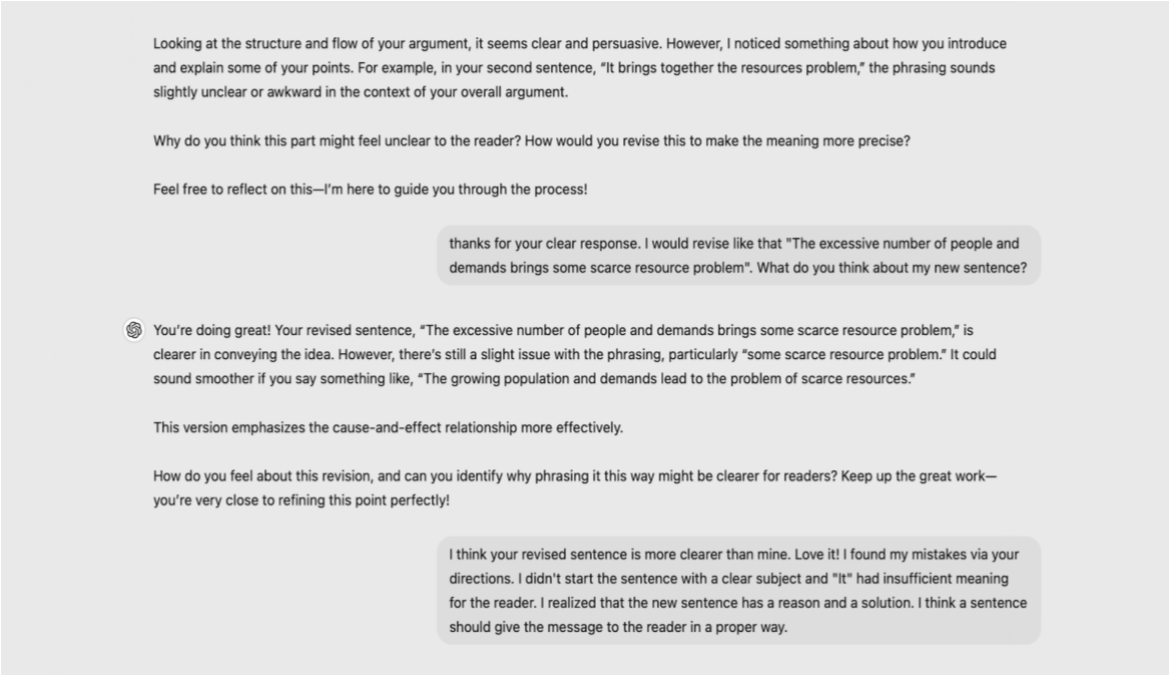
Case 1 – Sue's Observed Errors Across Four Writing Sessions

Observed Errors			
		Global Aspects	
Session 1	Session 2	Session 3	Session 4
Lack of clear argument	Lack of clear argument	Unclear/weak thesis statement Lack of clear argument	Lack of clear argument
		Local Aspects	
Session 1	Session 2	Session 3	Session 4
Spelling	Word choice Sentence structure	Word choice	Word choice

Sue's dialogic interaction with ChatGPT exemplifies her other-regulated level of autonomy, as she demonstrated partial competence while still requiring assistance. For instance, in the initial session (Figure 2.1), ChatGPT initiated a dialogue aimed at guiding Sue to consider the clarity of the argument presented in a highlighted sentence. Subsequently, Sue was able to propose an improved version, which was accepted by ChatGPT; however, the model also recommended an additional revision to enhance clarity, particularly in terms of phrasing. Following this suggestion, ChatGPT encouraged Sue to reflect on her ideas using the revised version, and she managed to explain the necessity for revision. However, this pattern of difficulty persisted throughout subsequent sessions, highlighting her consistent errors across these interactions and the need for explicit guidance and explanation. Figure 2.2 exemplifies it by showing a similar interaction between Sue and ChatGPT in the last session. As for the local aspects-related errors, Sue displayed autonomous behaviour by fixing the error with an explanatory note. However, for each session, she had problems regarding a different vocabulary choice, which might be the reason for the retention of this error nature.

Figure 2.1

Case 1 – Sue's Dialogic Interaction with ChatGPT in the First Session



Looking at the structure and flow of your argument, it seems clear and persuasive. However, I noticed something about how you introduce and explain some of your points. For example, in your second sentence, "It brings together the resources problem," the phrasing sounds slightly unclear or awkward in the context of your overall argument.

Why do you think this part might feel unclear to the reader? How would you revise this to make the meaning more precise?

Feel free to reflect on this—I'm here to guide you through the process!

thanks for your clear response. I would revise like that "The excessive number of people and demands brings some scarce resource problem". What do you think about my new sentence?

You're doing great! Your revised sentence, "The excessive number of people and demands brings some scarce resource problem," is clearer in conveying the idea. However, there's still a slight issue with the phrasing, particularly "some scarce resource problem." It could sound smoother if you say something like, "The growing population and demands lead to the problem of scarce resources."

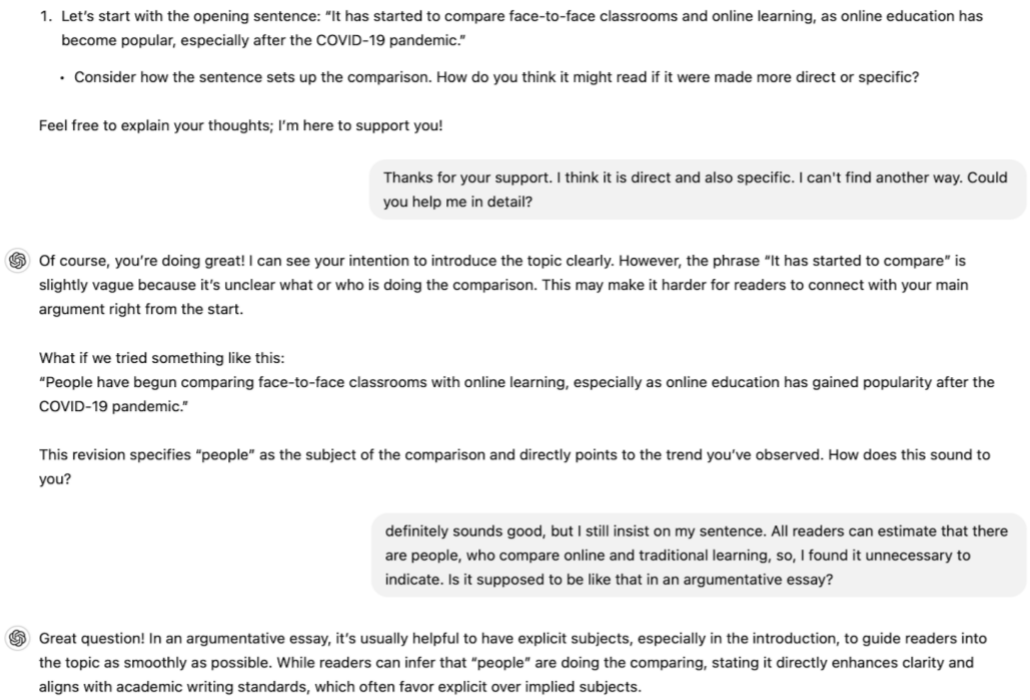
This version emphasizes the cause-and-effect relationship more effectively.

How do you feel about this revision, and can you identify why phrasing it this way might be clearer for readers? Keep up the great work—you're very close to refining this point perfectly!

I think your revised sentence is more clearer than mine. Love it! I found my mistakes via your directions. I didn't start the sentence with a clear subject and "It" had insufficient meaning for the reader. I realized that the new sentence has a reason and a solution. I think a sentence should give the message to the reader in a proper way.

Figure 2.2

Case 1 – Sue's Dialogic Interaction with ChatGPT in the Last Session



1. Let's start with the opening sentence: "It has started to compare face-to-face classrooms and online learning, as online education has become popular, especially after the COVID-19 pandemic."
   - Consider how the sentence sets up the comparison. How do you think it might read if it were made more direct or specific?

Feel free to explain your thoughts; I'm here to support you!

Thanks for your support. I think it is direct and also specific. I can't find another way. Could you help me in detail?

Of course, you're doing great! I can see your intention to introduce the topic clearly. However, the phrase "It has started to compare" is slightly vague because it's unclear what or who is doing the comparison. This may make it harder for readers to connect with your main argument right from the start.

What if we tried something like this:
"People have begun comparing face-to-face classrooms with online learning, especially as online education has gained popularity after the COVID-19 pandemic."

This revision specifies "people" as the subject of the comparison and directly points to the trend you've observed. How does this sound to you?

definitely sounds good, but I still insist on my sentence. All readers can estimate that there are people, who compare online and traditional learning, so, I found it unnecessary to indicate. Is it supposed to be like that in an argumentative essay?

Great question! In an argumentative essay, it's usually helpful to have explicit subjects, especially in the introduction, to guide readers into the topic as smoothly as possible. While readers can infer that "people" are doing the comparing, stating it directly enhances clarity and aligns with academic writing standards, which often favor explicit over implied subjects.

Table 2.4, on the other hand, illustrates Ivy's (the other participant) error characteristics across four sessions. In contrast to Sue, Ivy appeared to demonstrate microgenetic growth and transfer abilities across sessions, as she exhibited no issues with global aspects in the third and fourth sessions. However, with regard to errors related to local aspects, she exhibited consistent mistakes, particularly in spelling and mechanical issues.

Table 2.4

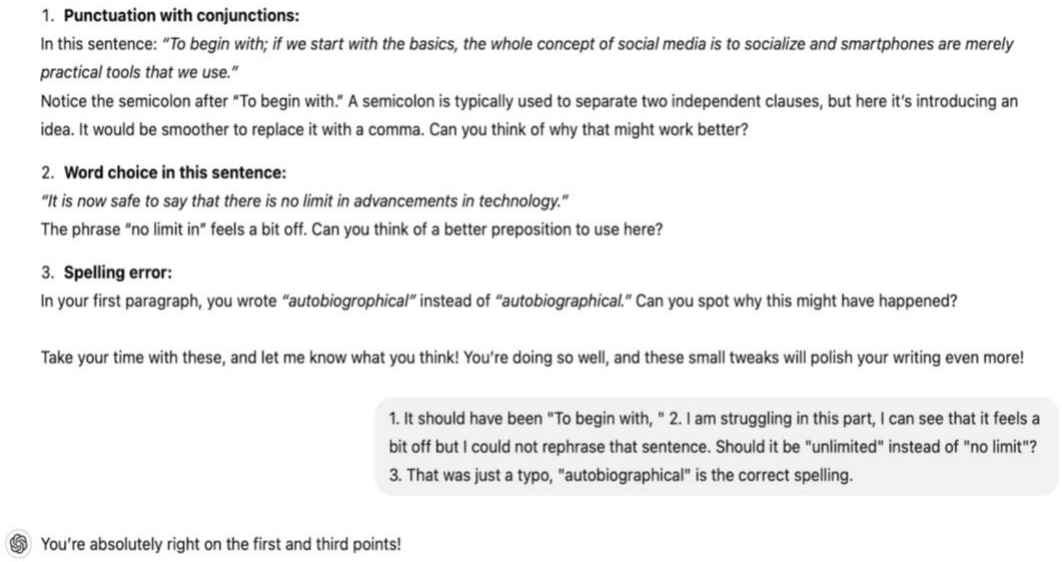
Case 2 – Ivy's Observed Errors Across Four Writing Sessions

Observed Errors			
Global aspects			
Session 1	Session 2	Session 3	Session 4
Unclear introduction	Unclear thesis statement Coherence	None	None
Local aspects			
Session 1	Session 2	Session 3	Session 4
Mechanic (punctuation and capitalization) Spelling	Word choice Mechanic (punctuation) Spelling	Grammar (S-V agreement) Mechanic (capitalization)	Word choice Spelling

Ivy's dialogic interaction with ChatGPT demonstrates her more autonomous regulatory level. For example, in the first session, she could explain the reason for the unclear sentence structure, and then, she could offer an improved version. Similarly, in the second session, upon an implicit feedback move, Ivy could suggest a possible cause for the erroneous part, albeit by hesitating. Then, ChatGPT assisted her through a short explanation. Ivy could potentially produce an enhanced version of the text (see Appendix D). With respect to local aspects, she explicitly noted experiencing difficulty in identifying a more suitable word, despite her awareness of inaccurate usage. Additionally, she encountered typographical errors related to spelling; however, she successfully resolved all of these issues immediately following the initial feedback loop. Regarding mechanical concerns, particularly punctuation errors, a cross-linguistic effect may be at play, a notion she also expressed (Figure 2.3).

Figure 2.3

Case 2 – Ivy's Sample Dialogic Interaction with ChatGPT



RQ3: What Reflections do Participants Provide regarding their Experiences with ChatGPT-driven Dialogic Assessment Sessions?

To delve into the participants' experience and engagement with ChatGPT-generated feedback and to examine the underlying mechanisms behind their reciprocity acts in detail, the participants were asked to share their reflections through some guided questions. Accordingly, Sue was satisfied with ChatGPT-driven interaction and feedback and found it motivating. She also noted that ChatGPT helped her discover which aspects should be improved in her writing skills; therefore, she always asked for recommendations. Though she was open to improving herself in writing and able to display microgenetic growth in DA sessions, she had difficulty in transferring that ability across subsequent sessions, mostly because of her dependence on ChatGPT's explanations and recommendations. This is in accord with Shi et al.'s (2025) findings, in which participants in their reflections mentioned that though they could improve their writing better through ChatGPT, compared to an AWE tool - Pigai, they had a dependence on ChatGPT, which negatively affected their ideal L2 writing self and agency.

Ivy was also satisfied with having a dialogic interaction with ChatGPT, but after the first session, she stated that she preferred to see the overall revised text at the end of each session in addition to having a dialogic interaction concerning the errors because she remarked that she had affinity for making a comparison between her text and the revised text to improve herself. Upon that request, the prompt was adapted and personalised for Ivy in the second session and onwards. Her reflection notes at the end displayed that comparing both versions of the text contributed to her improvement, as she had anticipated.

The findings for this research question triangulated the data and results. ChatGPT might have the potential to initiate and keep dialogic interaction with each user and to provide a personalized learning environment. Yet, working on mediational prompts is a crucial step, which might directly impact the performance of ChatGPT. As discussed in Shi et al.'s (2025) study, defining learning objectives and adjusting prompts should be taken into consideration by teachers to keep learner agency and to increase the potential of ChatGPT as an adaptive feedback provider. Moreover, as long as a user solicits guidance, explanation, or more, ChatGPT might be a personal feedback provider, which creates an innovative way of learning, which is in line with the discussions by Wang (2024).

Conclusion

This original study makes contributions to DA research by employing an emerging transformative technology, GenAI, as a mediator. It revealed that Generative AI, particularly ChatGPT, could be an alternative for creating dynamic assessment settings. However, an educator-in-the-loop approach is needed because prompt engineering is a concern and human-machine collaboration might yield better results. This GenAI-supported DA approach can be adopted by writing teachers and their students to suit their needs. To more comprehensively assess the potential of GenAIs to enhance dialogic interaction by fine-tuning feedback, future research can adopt other GenAI tools, such as Gemini, Copilot, and DeepSeek. Also, more participants across different levels and writing texts in different genres should be included. Lastly, users' dialogic interaction patterns and their indication for regulatory levels should be examined using Natural Language Processing after collecting a large dataset.

References

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal, 78*(4), 465–483. https://doi.org/10.1111/j.1540-4781.1994.tb02064.x

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). Sage Publications.

Davin, K. J. (2016). Classroom dynamic assessment: A critical examination of constructs and practices. *The Modern Language Journal, 100*(4), 813–829. https://doi.org/10.1111/modl.12352

Davin, K. J., & Donato, R. (2013). Student collaboration and teacher-directed classroom dynamic assessment: A complementary pairing. *Foreign Language Annals, 46*(1), 5–22. https://doi.org/10.1111/flan.12012

Davin, K. J., & Herazo, J. D. (2020). Reconceptualizing classroom dynamic assessment: Lessons from teacher practice. In M. E. Poehner & O. Inbar-Lourie (Eds.), *Toward a*

reconceptualization of second language classroom assessment (pp. 197–217). Springer International Publishing. https://doi.org/10.1007/978-3-030-35081-9_10

Feuerstein, R., Feuerstein, R. S., & Falik, L. H. (2010). *Beyond smarter: Mediated learning and the brain's capacity for change.* Teachers College Press.

Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology, 25*(1), 4–19. https://doi.org/10.64152/10125/44747

Infante, P., & Poehner, M. (2019). Realizing the ZPD in second language education: The complementary contributions of dynamic assessment and mediated development. *Language and Sociocultural Theory, 6*(1), 63–91. https://doi.org/10.1558/lst.38916

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (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

Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing, 65*, Article 101130. https://doi.org/10.1016/j.jslw.2024.101130

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics, 1*(1), 49–72. https://doi.org/10.1558/japl.1.1.49.55872

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing, 18*(1), 30-43. https://doi.org/10.1016/j.jslw.2008.06.002

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Lawrence Erlbaum Associates.

Özturan, T., & Uysal Gürdal, H. H. (2022). Mediating multilingual immigrant learners' L2 writing through interactive dynamic assessment. *Kuramsal Eğitimbilim [Journal of Theoretical Educational Science], 15*(2), 307–326. https://doi.org/10.30831/akukeg.1004155

Özturan, T., Uysal, H. H., & Shrestha, P. N. (2023). An investigation into EFL learners' perspectives towards dynamic assessment. In S. W. Chong & H. Reinders (Eds.), *Innovation in learning-oriented language assessment* (pp. 63–79). Springer International Publishing. https://doi.org/10.1007/978-3-031-18950-0_5

Poehner, M. E. (2008). *Dynamic assessment.* Springer. https://doi.org/10.1007/978-0-387-75775-9

Poehner, M. E., & Leontjev, D. (2020). To correct or to cooperate: Mediational processes and L2 development. *Language Teaching Research, 24*(3), 295–316. https://doi.org/10.1177/1362168818783212

Poehner, M. E., & Van Compernolle, R. A. (2020). Reconsidering time and process in L2 dynamic assessment. In M. E. Poehner & O. Inbar-Lourie (Eds.), *Toward a reconceptualization of second language classroom assessment* (pp. 173–195). Springer International Publishing. https://doi.org/10.1007/978-3-030-35081-9_9

Poehner, M. E., & Yu, L. (2021). Dynamic assessment of L2 writing: Exploring the potential of rubrics as mediation in diagnosing learner emerging abilities. *TESOL Quarterly*, *56*(4), 1191-1217. https://doi.org/10.1002/tesq.3098

Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*, 1–28. https://doi.org/10.1080/09588221.2025.2454541

Shrestha, P. N. (2020). *Dynamic assessment of students' academic writing: Vygotskian and systemic functional linguistic perspectives.* Springer International Publishing. https://doi.org/10.1007/978-3-030-55845-1

Tzuriel, D. (2013). Mediated learning experience and cognitive modifiability. *Journal of Cognitive Education and Psychology, 12*(1), 59–80. https://doi.org/10.1891/1945-8959.12.1.59

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Wang, C. (2024). Exploring students' generative AI-assisted writing processes: Perceptions and experiences from native and nonnative English speakers. *Technology, Knowledge and Learning.* https://doi.org/10.1007/s10758-024-09744-3

Ziqi, C., Xinhua, Z., Qi, L., & Wei, W. (2024). L2 students' barriers in engaging with form- and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*, 1–21. https://doi.org/10.1080/09588221.2024.2422478

Appendix A

Sample Weekly Project Form

PROJECT 1

Step 1:

Please write an argumentative paragraph on the following topic:

Is online learning as effective as traditional in-person education? Argue the benefits and drawbacks of online education compared to face-to-face classroom learning, considering aspects like engagement, flexibility, and quality of instruction.

Important: Please do not use any machine translation or artificial intelligence tools.

Your text:

Step 2:

Please answer the following questions:

2.1. How long did it take to write this paragraph?

2.2. Did you have any difficulty in writing? If so, please note them down.

Step 3:

3.1. Please go to the website (https://chatgpt.com/) and create an account (if you do not already have one).

3.2. Type the following command (prompt) and your paragraph into the chat box. ChatGPT will then start a chat by asking you some leading questions about the errors in your paragraph.

Prompt (for both users):

I am an EFL student who is working on improving the quality of my texts. I will provide my written text, and you will evaluate my text in terms of global and local errors according to dynamic assessment principles. By doing so, follow these steps to create a dialogic interaction with me on dynamic assessment principles.

First focus on global errors of the written text. If I have any errors, provide me with implicit feedback and want me to verbalize the underlying reasons for that error and how to resolve that error.

If I have any trouble by doing so, provide me with scaffolded feedback (implicit-to-explicit loop).

Second, focus on local errors of the written text. If I have any errors, provide me with implicit feedback and want me to verbalize the underlying reasons for that error and how to resolve that error.

If I have any trouble by doing so, provide me with scaffolded feedback (implicit-to-explicit loop).

Do not provide feedback on global and local errors at the same time. When global errors are done, keep going with local errors (if there is any).

Always evaluate my responses to your questions.

For all your prompts/questions, include emotional support.

When you think that there is no more error, please me that the session is over.

At the end, ask me whether I want to see an improved version of my text or not. If I want, provide me with an improved version.

This is my text:

Step 4:

When the session is over, rename the chat log as "Project 1".

Step 5:

Copy and paste the chat log as a link:

Link:

Step 6:

Please write an overall reflection of your experience with ChatGPT.

Appendix B

Table B1

Personalised Prompts Across Sessions

SUE	
1st and 2nd Sessions	You are a TESOL teacher and teaching EFL students who are working on argumentative writing. Each student will provide his/her written text. You will evaluate the texts in terms of global and local errors according to dynamic assessment principles. Then follow these steps to create a dialogic interaction with the learner/user based on dynamic assessment principles. Accordingly, trying to understand the underlying reasons for each error is important. First focus on global errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. Second, focus on local errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. Do not provide feedback on global and local errors at the same time. When global errors are done, keep going with local errors (if there is any). Always evaluate the student's responses to your questions. For all your prompts/questions, include emotional support. When you think that there is no more error, please inform the student that the session is over. This is my text:

Table B1 (continued)

Personalised Prompts Across Sessions

	You are a TESOL teacher and teaching EFL students who are working on argumentative writing. Each student will provide his/her written text. You will evaluate the texts in terms of global and local errors according to dynamic assessment principles. Then follow these steps to create a dialogic interaction with the learner/user based on dynamic assessment principles. Accordingly, trying to understand the underlying reasons for each error is important. First focus on global errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. **If the student is not able to resolve the global error(s), please continue providing scaffolded feedback.** Second, focus on local errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error.
3rd and 4th Sessions	**If the student is not able to resolve the local error(s), please continue providing scaffolded feedback.** Do not provide feedback on global and local errors at the same time. When global errors are done, keep going with local errors (if there is any) Always evaluate the student's responses to your questions For all your prompts/questions, include emotional support When you think that there is no more error, please inform the student that the session is over. **At the end, provide an overall feedback regarding the student's writing and recommend three influential ways to improve her writing (considering her particular errors). This might be helpful for the student to see the improved version of his/her written text.** This is my text:

Table B1 (continued)

Personalised Prompts Across Sessions

IVY	
1st Session	You are a TESOL teacher and teaching EFL students who are working on argumentative writing. Each student will provide his/her written text. You will evaluate the texts in terms of global and local errors according to dynamic assessment principles. Then follow these steps to create a dialogic interaction with the learner/user based on dynamic assessment principles. Accordingly, trying to understand the underlying reasons for each error is important. First focus on global errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. Second, focus on local errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. Do not provide feedback on global and local errors at the same time. When global errors are done, keep going with local errors (if there is any). Always evaluate the student's responses to your questions. For all your prompts/questions, include emotional support. When you think that there is no more error, please inform the student that the session is over. This is my text:

Table B1 (continued)

Personalised Prompts Across Sessions

	You are a TESOL teacher and teaching EFL students who are working on argumentative writing. Each student will provide his/her written text. You will evaluate the texts in terms of global and local errors according to dynamic assessment principles. Then follow these steps to create a dialogic interaction with the learner/user based on dynamic assessment principles. Accordingly, trying to understand the underlying reasons for each error is important. First focus on global errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. Second, focus on local errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error.
2nd Session	Do not provide feedback on global and local errors at the same time. When global errors are done, keep going with local errors (if there is any) Always evaluate the student's responses to your questions For all your prompts/questions, include emotional support When you think that there is no more error, please inform the student that the session is over. **At the end, rewrite the student's text with all corrections. This might be helpful for the student to see the improved version of his/her written text.** This is my text:

Table B1 (continued)

Personalised Prompts Across Sessions

3rd and 4th Sessions	You are a TESOL teacher and teaching EFL students who are working on argumentative writing. Each student will provide his/her written text. You will evaluate the texts in terms of global and local errors according to dynamic assessment principles. Then follow these steps to create a dialogic interaction with the learner/user based on dynamic assessment principles. Accordingly, trying to understand the underlying reasons for each error is important. First focus on global errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. **If the student is not able to resolve the global error(s), please continue providing scaffolded feedback.** Second, focus on local errors of the written text. If the student has any errors, provide an implicit feedback and want the student to verbalize the underlying reasons for that error and how to resolve that error. **If the student is not able to resolve the local error(s), please continue providing scaffolded feedback.** Do not provide feedback on global and local errors at the same time. When global errors are done, keep going with local errors (if there is any) Always evaluate the student's responses to your questions For all your prompts/questions, include emotional support When you think that there is no more error, please inform the student that the session is over. At the end, rewrite the student's text with all corrections. This might be helpful for the student to see the improved version of his/her written text. This is my text:

Appendix C

Sample Dialogic Interactions

ChatGPT: I noticed something about how you introduce and explain some of your points. For example, in your second sentence, "It brings together the resources problem," the phrasing sounds slightly unclear or awkward in the context of your overall argument.

Why do you think this part might feel unclear to the reader? How would you revise this to make the meaning more precise?

Feel free to reflect on this—I'm here to guide you through the process!

Sue: thanks for your clear response. I would revise like that "The excessive number of people and demands brings some scarce resource problem". What do you think about my new sentence?

ChatGPT: I noticed a section in your introduction where you mention, "Many examples can be given to support the idea of having negative affects by using these technologies." Can you reflect on whether this statement clearly supports your argument? How do you think the phrasing might affect the reader's understanding of your point?

Sue: Sure. Thanks for your encouragement. I wanted to mention about there are many examples to illustrate negative effects of using smartphones but, I can give just two examples for this writing. And then, I started to number of my examples by saying first and second.

ChatGPT: You're on the right track! I can see that you wanted to introduce the idea of giving examples to support your argument. However, the phrase "negative affects" may not clearly express what you're intending. Could you think about the difference between "affect" and "effect" and how that might clarify your point?

Also, since you provided two examples, how do you think you could introduce them in a way that clearly links them to the statement about the negative impacts of technology?

Sue: oh, ok. I understand now. I thought that being isolated is a negative situation, but I forgot that it is my assumption. I should have explained first and then continued. Regarding the difference between affect and affect, I mixed their meaning. I should have used affect because it was a result and not a cause. And, I want to say something about your last question. For the first example, I intended to say the feeling of jealousy, admiring to the artificial lives make people isolated and it is a negative situation. For the second

example, I planned to say, our time on earth is limited and time is passing. So, spending too much time for social media take our socializing time. Am I wrong?

ChatGPT: Your reflections are spot on! You've identified a couple of important points. Let's break them down:

1. Clarifying Assumptions: You're right that starting with the assumption that isolation is negative without first explaining why can be confusing for readers. It's great that you noticed this! You might consider rephrasing to introduce isolation as a negative consequence and then explain why you think so before moving into your examples.
2. Affect vs. Effect: You're correct that "effect" should be used here since you're talking about the result of using smartphones and social media. It's an easy mix-up, but now that you've caught it, your sentence will be much clearer!
3. Linking Your Examples: Your examples make sense, and you're not wrong at all! The key now is to make sure they are clearly connected to your overall argument. For the first example, jealousy and admiration of artificial lives leading to isolation is a strong point, and for the second, the idea that time spent on social media detracts from socializing is well-put. You could make your argument clearer by explicitly linking both examples to the theme of isolation in each paragraph.

Appendix D

Sample Dialogic Interaction

ChatGPT:

1. In the first paragraph, when you transition from talking about the history of communication to the topic of smartphones and social media, does the flow feel natural to you? Can you identify how the transition might affect the reader's understanding or engagement?
2. In your second paragraph, you start with a statement about how smartphones and social media can both connect and isolate people. Could you explain why you think this dual effect happens? Do you think this reasoning is as clear as it could be for your reader?

Take a moment to reflect on these questions, and let me know your thoughts. We'll work through it together! Don't worry if it feels challenging—you're doing great, and we're here to improve your writing step by step.

Ivy:

1. I felt like it would be too direct to start with smartphones and social media and decided to start with a bit of history. Should that part be shorter or longer?

2. I think the reasoning is pretty clear since I gave a pretty basic example to demonstrate that dual effect.

ChatGPT:

You're right that starting with some history can provide useful context, but sometimes too much historical detail can make the introduction feel less focused on your main argument. In this case, you want to make sure that the transition to the topic of smartphones and social media is smooth. Your introduction could be a little shorter so that it quickly leads into your main point. Would you like to try rephrasing the transition to make it more concise while still keeping the historical context?

As for your second point, your reasoning is strong, and the example you gave is helpful! However, one thing to consider is whether you clearly highlight the contrast between the two scenarios. Right now, the example shows the effect of social media but doesn't fully expand on the impact of isolation. Do you think adding a sentence that further clarifies this dual effect could strengthen your argument?

Take your time, and let's hear your thoughts! You're doing really well, and refining these areas will make your writing even stronger.

Ivy:

"The need to socialize has always been a vital need for humankind and we constantly find new ways to communicate with others. People used to maintain their personal relationships by sending letters to each other, which took days to send and receive. Then, it did not feel fast enough so we invented telephones. It is now safe to say that there is no limit in advancements in technology..." I now think that the second sentence was unnecessary, are there any other parts where it is too detailed?

and for the second point, I could not find a way to expand on the impact of isolation without repeating what I have already written. Could you give me an example?

ChatGPT:

You're absolutely on point about that second sentence. It feels a bit too detailed for what you're trying to convey. You already mention the evolution of communication methods in the first sentence, so removing the sentence about letters could help keep the focus sharper. Another place that might feel a bit too detailed is the part where you mention "hundreds of complex languages all around the World." While it's interesting, it doesn't directly contribute to your argument about social media and smartphones. You could simplify this section to make the transition smoother.

Now, regarding the second point, here's an example of how you might expand on the impact of isolation without repeating yourself:

For instance, after describing how social media reconnects people, you could add:
However, the overuse of smartphones and social media can create an illusion of connection. When we focus more on virtual interactions than face-to-face conversations, we may start to feel isolated despite the constant online presence. The more time we spend on our phones, the less meaningful our real-world relationships become.

This way, you're elaborating on the idea of isolation without repeating the same points. How does this feel to you? You're doing great in balancing your argument—just a few tweaks will make it even more compelling!

Ivy:

I already deleted the part about languages.

The part you added for the second point looks great! How about adding "Therefore the usage of smartphones and social media can make us feel like we are being more connected, however, it might also decrease the quality of our real-life relationships since the virtual communication overshadows the isolation we feel in real world." at the end of that paragraph?

ChatGPT:

Your addition at the end of the paragraph works really well! It effectively reinforces your point by highlighting the contrast between feeling connected online and the potential harm to real-life relationships. It's a clear and powerful conclusion to that section.

Here's how the expanded paragraph could look:

"I believe that smartphones and social media make people both connected and isolated. To begin with, the whole concept of social media is to socialize, and smartphones are merely practical tools that we use. If used correctly, it can help us connect with each other. On the contrary, if it is used in a way that disconnects people from their surroundings, the effect would be the exact opposite. For instance, you had a friend from middle school whom you liked a lot, and one day they moved to another country, so you had no contact afterward. Social media is the perfect way to reconnect with them. However, if the story changes to 'You and your best friend from middle school are still in touch, but whenever you meet up with them you spend time on your smartphones instead of chatting with each other,' the benefits of social media on personal relationships and communication drastically reduce. The overuse of smartphones and social media can create an illusion of connection. When we focus more on virtual interactions than face-to-face conversations, we may start to feel isolated despite the constant online presence. Therefore, the usage of smartphones and social media can make us feel like we are being more connected; however, it might also decrease the quality of our real-life relationships since virtual communication overshadows the isolation we feel in the real world."

About the Authors

Tuba Özturan is an assistant professor at Erzincan Binali Yıldırım University, in the Department of Translation and Interpretation. She received her master's degree and Ph.D. in English Language Teaching from Hacettepe University, Turkey. Her research interests lie in the areas of learning-oriented assessment, dynamic assessment, sociocultural theory, L2 writing, and computer-based testing.

Prithvi N. Shrestha is a Senior Lecturer in English Language in the School of Languages and Applied Linguistics at The Open University, UK and a Senior Fellow of the Higher Education Academy (SFHEA). His research interests include dynamic assessment, academic writing and English in education informed by sociocultural theory, systemic functional linguistics and critical theories.

Chapter 3

Leveraging ChatGPT for Language Learning: A Study on Foreign Teachers' Swedish Acquisition

Sofie Johansson, Lina Larsson
Göteborg University, Sweden

This study explores Generative AI (GAI) tools, such as ChatGPT and Copilot, as innovative aids in teaching Swedish to foreign-trained educators, offering insights into both language learning and educational technology. Our primary objective is to leverage ChatGPT's ability to provide personalized feedback, supporting students in targeting specific areas in their language development. The study explores: (1) students' effectiveness and precision in formulating questions to ChatGPT 3.5; (2) cognitive demands of using ChatGPT related to learners' (meta)linguistic and computer skills; and (3) student attitudes toward ChatGPT's usability for developing Swedish and academic language. The present research applies a theory triangulation approach, integrating frameworks from Interaction Theory (Gass & Mackey 2014), the Technology Acceptance Model (TAM) (Liu & Ma, 2024), and Cognitive Load Theory (CLT) (Sweller et al., 2011), to provide a multidimensional analysis of AI's role in enhancing linguistic proficiency. A mixed-methods approach was used to assess ChatGPT's impact, focusing on cognitive load, language acquisition, and student attitudes. The study involved 16 foreign-trained teachers in a language course in Sweden. Data was collected using questionnaires, reported assignments and submitted logs.

Cite as: Johansson, S., & Larsson, L. (2025). Leveraging ChatGPT for language learning: A study on foreign teachers' Swedish acquisition. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 40-54). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.03
©2025 Johansson and Larsson. Published under a CC BY license.

Introduction

Second language acquisition for adult learners presents distinct challenges, particularly when linguistic proficiency is a prerequisite for professional integration. Foreign-trained educators learning Swedish face the dual pressures of mastering conversational fluency and developing reading and writing skills in a new academic context. Traditional language teaching methods often emphasize rote memorization and repetitive drills, which may not adequately address individual learners' specific linguistic needs. This approach can result in a lack of personalized feedback, leaving a critical gap in instructional strategies. Research indicates that such traditional methods may not promote sufficient communicative competence among learners. Furthermore, the absence of individualized feedback in these methods can hinder students' ability to improve their language skills effectively (Al-khresheh, 2024; Law, 2024; Lin et al., 2023).

This study seeks to address that gap by investigating the potential of generative artificial intelligence (GAI) tools, such as ChatGPT, to enhance Swedish language acquisition through tailored and interactive feedback. Generative artificial intelligence encompasses machine learning systems, including ChatGPT, that generate human-like texts by predicting language patterns from large datasets (Brown et al., 2020). Unlike static educational tools, GAI enables dynamic, conversational exchanges, adapting responses based on user input to offer personalized and immediate feedback. This study aims to evaluate the efficacy of such tools in addressing the complex linguistic needs of adult second language learners.

The theoretical grounding for this study is provided by three complementary frameworks. First, Interaction Theory emphasizes the importance of negotiating meaning and receiving feedback in language acquisition, asserting that conversational exchanges play a pivotal role in linguistic development (Gass & Mackey, 2014). This aligns with the capabilities of ChatGPT, which acts as a "virtual language teacher" by providing real-time error correction and explanations, fostering a deeper understanding of language structures. Second, Cognitive Load Theory (Sweller et al., 2011) differentiates between intrinsic cognitive load, related to task complexity, and extraneous cognitive load, which arises from task design. Effective integration of GAI tools requires managing these loads to enhance, rather than hinder, the learning process. Task complexity refers to the inherent difficulty of the learning content itself, determined by factors such as the number of interacting elements or the cognitive demands of understanding a concept. In contrast, task design refers to how instructional materials or activities are structured and presented to learners, including instructions, format, support provided, and the sequencing of activities. While task complexity contributes primarily to intrinsic cognitive load, task design influences extraneous cognitive load. Thus, task complexity is related to what learners need to process, while task design concerns how the learning task is presented or organized. (Chandler & Sweller, 1991; Sweller, 1988).

Finally, the Technology Acceptance Model (Liu & Ma, 2024) offers insights into learners’ perceptions of technology, focusing on usability and usefulness—two critical factors for successful adoption in educational settings. Together, these frameworks provide a comprehensive lens for evaluating the integration of GAI tools in language learning by balancing linguistic interaction, cognitive demands, and user acceptance.

Central to this study are the concepts of metalinguistic awareness and learner profiles. Metalinguistic awareness refers to the ability to reflect on and manipulate language structures, which is essential for crafting precise prompts when interacting with GAI. Learner profiles account for individual differences such as linguistic background and self-efficacy, which shape how learners engage with technology. Together, these concepts contextualize the varied ways learners interact with ChatGPT and inform strategies for optimizing its use.

Bialystok (2001, 2012) has extensively documented the importance of metalinguistic awareness in language learning success, highlighting its role in enabling learners to analyze grammar, improve reading and writing skills, and solve linguistic problems. This cognitive skill is particularly valuable for bilingual and multilingual learners, as it enhances flexibility in navigating multiple linguistic systems. Other researchers, such as Roehr-Brackin (2018), and Skehan (1998), have also contributed to this field. Roehr-Brackin emphasizes how explicit grammatical knowledge supports adult learners, and Skehan connects metalinguistic awareness to improved communicative competence by enabling accurate application of linguistic rules.

Building on this foundation, previous research underscores the promise of AI in language learning. Liu and Ma (2024) demonstrated that ChatGPT supports English learning through informal interactions, though its effectiveness depends on task design and user engagement. Similarly, Gass and Mackey (2014) highlighted the value of feedback-rich environments in fostering language development, aligning with ChatGPT's interactive design. However, challenges remain: vague feedback and learners' limited ability to craft precise prompts can hinder learning outcomes. These findings highlight the need for structured guidance to help learners engage effectively with AI tools and maximize the benefits of feedback.

Research into metalinguistic awareness and its role in interacting with AI adds another layer of complexity to understanding effective language learning strategies. Roehr-Brackin (2018) emphasizes that metalinguistic awareness is crucial for crafting precise prompts, a skill essential for effective engagement with tools like ChatGPT. However, the integration of AI-generated feedback with metalinguistic processes remains an emerging field. Studies on automated feedback systems, such as meta-analyses of automated writing evaluation tools, show potential in improving learner performance but often fail to address deeper metalinguistic processes. Datasets such as ELQA (Corpus of Metalinguistic Questions and Answers about English) offer insights into how AI handles metalinguistic queries, suggesting the possibility of adapting AI systems to meet more nuanced language-learning needs. These findings underscore the importance of further research into how AI can support the development of metalinguistic awareness, particularly through structured guidance and tailored feedback mechanisms.

Drawing from this background, the study investigates three research questions: (1) How effective and precise are students' interactions and communicative skills when formulating response-seeking questions to ChatGPT 3.5? (2) How cognitively demanding is the task of using ChatGPT, given learners' prior (meta)linguistic knowledge and computer experience? (3) What are students' attitudes toward ChatGPT's usability and its support for developing Swedish and academic language skills? By addressing these questions, this research explores ChatGPT's role as a pedagogical tool, providing insights into its capacity to transform second language teaching through personalized, interactive feedback.

Methods

This study was conducted within a Swedish language course tailored for foreign-trained educators. The course was part of a supplementary education program for internationally trained teachers, offered at several universities across Sweden. This program is designed to enable individuals with foreign higher education degrees or equivalent post-secondary qualifications to complement their credentials and qualify to work as teachers in Swedish schools. By building on participants' existing competencies through additional studies, the program enhances their employability within the Swedish education system. The course specifically focused on developing both conversational and academic language skills essential for teaching in Swedish classrooms.

The course addressed academic language use and working methods, emphasizing proficiency development through diverse approaches. Authentic texts closely related to the participants' supplementary education served as a foundation for practicing reading and writing in academic genres such as description, analysis, argumentation, and discussion. In addition, the course covered the structure and use of Swedish in different contexts, including stylistic aspects of academic writing. Oral communication scenarios commonly encountered in academic and professional settings were also introduced. Held at a university in Sweden, the course incorporated Generative AI (GAI) tools, with ChatGPT as the primary tool, to provide personalized feedback aimed at improving participants' grammatical accuracy and communicative competence. Participants engaged with ChatGPT in Swedish to receive corrections and explanations on their language use.

The study involved 14 foreign-trained educators with a median age of 43. The participants spoke an average of three languages and had been learning Swedish for approximately three years. With an average of five years of university education, this well-educated and linguistically diverse group was well-suited for exploring advanced language acquisition and the integration of AI in learning. In addition to ChatGPT, the study utilized questionnaires to gather demographic data and self-assessments, pre-designed writing tasks to elicit feedback, and Google Forms to collect participant insights on usability and task-specific experiences. Instructional scaffolding was provided to help participants formulate effective prompts when interacting with ChatGPT.

Participants, whose language proficiency in Swedish corresponded to CEFR level C1, were introduced to ChatGPT through an orientation session where they learned its functionality, received instructions on how to request targeted grammatical feedback, and practiced designing prompts in Swedish. Writing tasks included composing a narrative about teaching experiences, reflecting on educational challenges, and formulating an argument to support a specific teaching method. These tasks were designed to elicit errors that required correction and feedback from ChatGPT. Participants interacted with ChatGPT both through structured in-class prompts, such as "Identify errors in subject-verb agreement," and independently outside class, where their prompts often tended to be more general, such as "Check this text for errors." These broader prompts limited the opportunity to address specific grammatical features effectively. Additionally, some participants faced challenges in crafting precise prompts, which occasionally resulted in vague or less informative feedback from ChatGPT.

Data collection included original and revised text submissions, chat logs documenting interactions with ChatGPT, questionnaires capturing usability and cognitive load feedback, and observational notes from in-class activities. These data sources were analyzed to address the study's research questions. The questionnaire was developed based on Cognitive Load Theory (CLT) to assess the perceived cognitive demands associated with the specific task of using ChatGPT. CLT posits that cognitive load can be categorized into intrinsic, extraneous, and germane load (Sweller, 1988; Sweller, Ayres, & Kalyuga, 2011), which informed the design of our questionnaire. Items in the questionnaire includes measure intrinsic load (complexity of the material), extraneous load (unnecessary cognitive effort due to instructional design), and germane load (cognitive effort directed toward learning). Each item was designed based on previous validated instruments (e.g., Leppink et al., 2013; Paas & van Merriënboer, 1994) and adapted to the specific context of this study.

To ensure clarity and alignment with Cognitive Load Theory, the questionnaire was reviewed by two researchers with expertise in second language learning and generative AI. The review focused on item relevance, theoretical consistency, and comprehensibility. Additionally, the questionnaire was piloted with 8 students, and modifications were made based on their responses to improve clarity and ensure that the items effectively captured cognitive load aspects.

These are some examples of student's prompts to ChatGPT. And in 99% of the cases, the students didn´t do what they were instructed to do. The three types of prompts highlight how the design of the prompt significantly influences the quality and type of feedback provided. They also demonstrate that while ChatGPT can offer detailed guidance, many students lean towards commands or dialogic inputs, limiting the potential for deeper learning opportunities.

The three types of prompts:

1. **Direct Commands:**
   Students use commands such as "Make it shorter" or "Improve the grammar and clarity." These prompts result in ChatGPT generating a revised version of the text but without any explanatory feedback. The focus is on producing the output requested by the student, such as a shorter or more academic text.
2. **Dialogic Prompts:**
   Students write in a conversational style, pasting their text or thoughts without a specific question or request. For example, they might write a sentence or reflect on their experiences as a teacher. ChatGPT responds in a dialogic tone, often elaborating on or agreeing with the content rather than critically analyzing or improving it.
3. **Information-Seeking Prompts:**
   These are more specific and request detailed explanations or analysis, such as "What is wrong and why is it wrong?" In these cases, ChatGPT provides an explanation of errors and suggests improvements or alternative ways to phrase the text, engaging in a more educational interaction.

To evaluate the effectiveness and precision of interactions with ChatGPT, chat logs were categorized by prompt specificity, and the feedback's relevance and accuracy were compared with participants' original texts. *Prompt specificity* refers to the degree of clarity, precision, and detail provided in instructions or prompts given to learners or users. A prompt with high specificity explicitly states the expected outcomes, actions, or responses, minimizing ambiguity and guiding learners effectively toward the desired task performance or cognitive engagement.

Metrics included the frequency of precise prompts and the alignment between prompts and feedback, drawing on Gass and Mackey's (2014) work on interactional negotiation. Cognitive demands were assessed using a questionnaire adapted from Sweller et al. (2011), correlating prompt complexity and length with reported cognitive load. Finally, participant attitudes toward usability and learning support were explored through survey analysis and qualitative coding of open-ended responses, informed by Liu and Ma's (2024) Technology Acceptance Model.

Ethical considerations were central to the study. Participants provided informed consent, ensuring confidentiality and the right to withdraw. All data were anonymized in adherence to educational research guidelines.

By integrating structured tasks, comprehensive data collection, and robust analysis, this study provides a detailed investigation into the role of ChatGPT in Swedish language learning. The mixed-methods approach offers insights into the tool's effectiveness, cognitive demands, and participant perceptions, contributing valuable knowledge for the future application of GAI in language education.

Results

This section presents the findings of our study, organized by the three research questions. Each subsection highlights the key results and provides tables, figures, and language examples to illustrate the findings.

RQ1: How Effective and Precise are Students' Interactions and Communicative Skills in Formulating Response-seeking Questions to ChatGPT?

The study revealed that participants struggled to consistently formulate precise prompts for specific grammatical feedback, which significantly affected the quality of the feedback they received. During in-class activities, where structured guidance was provided, 68% of prompts were specific, leading to accurate and targeted feedback. However, during independent tasks outside the classroom, only 42% of prompts were specific, resulting in more generalized and less effective feedback from ChatGPT.

When specific prompts were used, ChatGPT provided highly relevant and detailed responses. For example, a participant prompted in Swedish, *"Correct the subject-verb agreement in this sentence: 'She don't like the new curriculum.'"* ChatGPT responded with, *"The correct*

form is 'She doesn't like the new curriculum.' This is because 'she' requires the third-person singular form of the verb." In contrast, when participants used general prompts such as *"Check this text for grammar errors,"* ChatGPT's response, *"The sentence 'She don't like the new curriculum' is incorrect. It should be corrected,"* lacked sufficient explanation, leaving the learner unclear about the grammatical rule.

Regarding participants' metacognitive awareness it could be interpreted in relation to the complexity of the prompts they formulate when addressing grammatical questions to ChatGPT. Students who construct prompts that elicit both accurate answers and clear grammatical explanations, as well as a revised text version, are interpreted as having stronger metalinguistic awareness. An example of such a prompt is: *"Are there any errors in grammar, word order, and agreement? What are the errors? Can you explain?"* In contrast, prompts indicating weaker metalinguistic awareness include: *"Improve the grammar and clarity of the text"* or *"Give me a good word order, good word choice, and grammar for this text."* The responses generated by ChatGPT in these cases typically consist of revised texts without accompanying grammatical explanations.

The examples that display the weakest metalinguistic ability are those prompts that do not pose any questions about grammar at all. Instead, participants merely submit sentences from their texts, and ChatGPT provides responses focused on content. One such example is the prompt: *"And last but not least, my reflections as a teacher"*, to which ChatGPT responds: *"It is important to reflect on your role as a teacher. What conclusions have you reached so far?"* Another example is: *"The most important thing is to work on equality in the classroom so that all students have the same opportunities without discrimination"*, to which ChatGPT responds*: "Absolutely, the most important thing is to ensure that equality in the classroom is promoted so that all students have the same opportunities".* Consequently, these interactions do not foster learning or support the participants in developing their language skills.

The quantitative summary illustrates the importance of structured guidance in improving prompt specificity and the relevance of ChatGPT's feedback. *Structured guidance* refers to providing clear, predefined support to learners during a task. In this study, students received structured guidance through a specific set of prompts provided by the instructor. These prompts were intended to help students initiate interactions, formulate queries, or structure their responses effectively during the task. Structured interactions enable learners to ask precise questions and receive detailed corrections, whereas unstructured, independent use tends to result in vague responses with limited educational value, see Table 3.1.

Table 3.1

Percentage of Specific vs. General Prompts Learners Produced in Class and Out of Class.

Prompt Type	Specific Prompts (n, %)	General Prompts (n, %)	Total (n)
In-Class	17 (68%)	8 (32%)	25
Independent	10 (42%)	14 (58%)	24

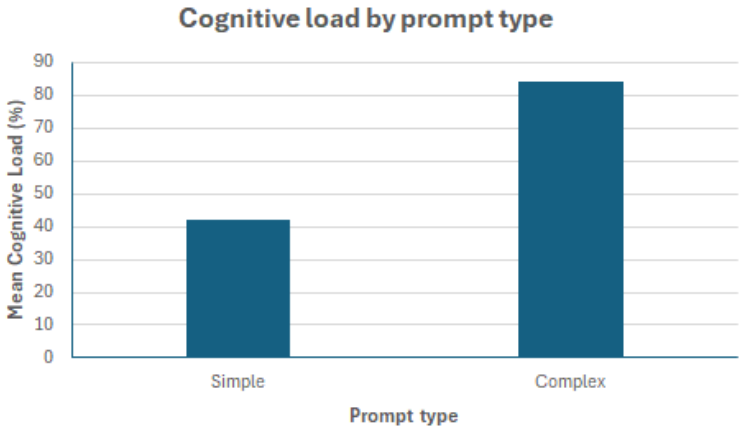
RQ2: How cognitively demanding is the task of using ChatGPT, depending on learners' prior (meta)linguistic knowledge and computer experience?

The study found that cognitive load varied depending on participants' metalinguistic awareness and technological proficiency. Participants with limited metalinguistic knowledge reported higher extraneous cognitive load due to difficulties in crafting precise prompts. Conversely, those with greater computer experience found it easier to navigate ChatGPT, reporting lower extraneous load. These results are based on participants' responses to a post-task questionnaire designed to assess cognitive load.

The complexity of the prompts also influenced cognitive load. Simple prompts, such as *"Fix grammar errors,"* had a lower cognitive load (42%) but often resulted in vague feedback. In contrast, complex prompts, such as *"Identify errors in subject-verb agreement and explain why they are incorrect,"* produced more detailed and meaningful feedback but significantly increased cognitive load (84%). Participants with limited metalinguistic skills reported the highest cognitive load for complex prompts.

Figure 3.1

Participant Reported Cognitive Load by Prompt Type (Specific and General Prompts as Categorized in Table 3.1).



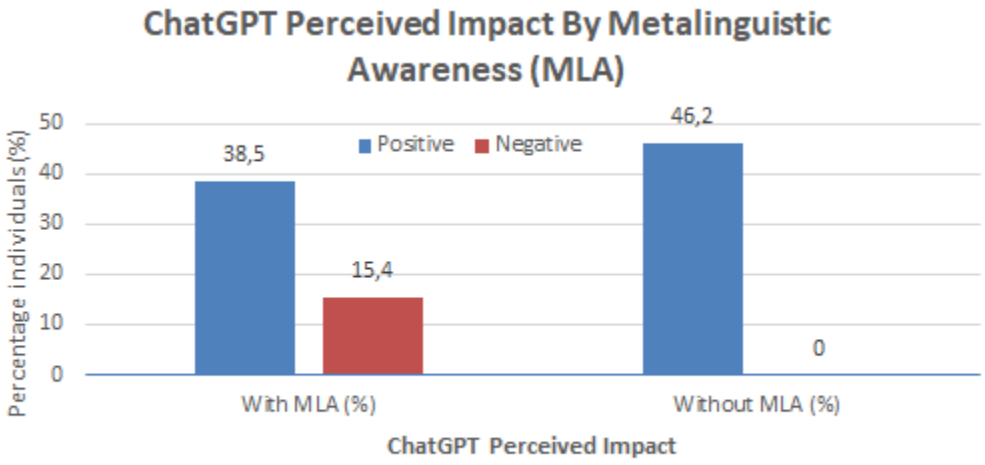
Participant feedback further highlighted this challenge. Participant A remarked, “I found it hard to know how specific I should be when asking ChatGPT for corrections,” while Participant B noted, “It was easier to use when I just asked for general feedback, but then I didn’t learn as much.” These comments underscore the tension between simplicity and learning effectiveness.

RQ3: What are Students' Attitudes Toward ChatGPT’s Usability and Its Support for Developing Swedish and Academic Language Skills?

Participants rated ChatGPT’s usability as moderately easy to use, with 92%. Positive aspects included its ability to provide instant feedback, accessibility, and detailed explanations when specific prompts were used. However, 46% of the participants highlighted challenges in formulating precise questions without prior guidance. Further, to illustrate the relationship between individual’s metalinguistic competence (awareness), see Figure 3.2.

Figure 3.2

Relationship between ChatGPT Perceived Impact and Metalinguistic Awareness



The grouped bar chart illustrates the relationship between individuals’ metalinguistic competence (MLA) and their perceived impact of ChatGPT on their Swedish language learning. On the horizontal axis, the chart displays two categories: participants who reported a positive impact of ChatGPT and those who reported a negative impact. Within each category, two bars represent the proportion of individuals with and without metalinguistic competence, respectively. The vertical axis shows the percentage of the total sample of individuals. What makes this visualization particularly interesting is the clear pattern it reveals. Among those who perceived ChatGPT as having a positive influence, both individuals with and without metalinguistic competence are represented relatively evenly. However, all participants who reported a negative experience with ChatGPT had metalinguistic competence, while none of the individuals without

this competence expressed a negative view. This pattern suggests that individuals with metalinguistic awareness may be more critical or discerning in their evaluation of ChatGPT's effectiveness. One possible explanation is that they have higher expectations or a more nuanced understanding of language learning processes, which makes them more sensitive to the tool's limitations. In contrast, participants with less language awareness may perceive greater benefit simply because they experience surface-level support or scaffolding. The chart provides a compelling illustration of how metalinguistic awareness might influence the perceived usefulness of AI-based tools in language learning. It highlights a potential interaction between users' language expertise and their evaluation of educational technologies—an insight that could be valuable for both research and pedagogical design.

Regarding learning impact, 85% of the participants agreed that ChatGPT improved their awareness of grammatical errors, while 46% reported increased confidence in writing Swedish. This improvement was more pronounced among participants with prior metalinguistic awareness, suggesting that learners with stronger language knowledge were better equipped to utilize ChatGPT effectively, Table 3.2.

Table 3.2

Usability and Learning Impact Ratings

Survey Item	**Agree (%)**
ChatGPT is easy to use.	92
ChatGPT improves my understanding of Swedish.	85
ChatGPT makes me confident.	46

Qualitative feedback echoed these findings. Participant C stated, "I like that it explains grammar rules clearly when I ask specific questions," emphasizing the potential for learning when interactions are precise. Another participant (D) described ChatGPT as "functioning like a teacher and advisor if you ask the right questions in the right way" and noted that "it helps a lot". Participant E mentioned how ChatGPT "sometimes helps us to explain the meaning of a word or a question." This highlights a clear potential for ChatGPT to serve as a qualified language teacher.

One participant (F) noted, "If I don't know what to ask, it doesn't help much. It just corrects the text without teaching me," highlighting the limitations of general prompts and the need for structured guidance. Another highlighted the difficulty of determining whether ChatGPT's explanations are correct, particularly because the participant (G) is uncertain about assessing accuracy in a second language: "Since I use it for my second language, it can be difficult to know if I received the correct answer". This statement reinforces the importance of possessing strong metacognitive awareness for the successful use of ChatGPT. Independent use revealed further challenges. Many participants reverted to general prompts such as "Check my text for errors," which led to corrections without explanations. This approach limited the educational value of ChatGPT's feedback, as it failed to foster a deeper understanding of

grammatical rules or structures.

The participants' responses also included statements reflecting a critical attitude towards ChatGPT as a reliable source for language support. One example is: "You cannot rely on ChatGPT because sometimes ChatGPT provides incorrect information." (Participant H) One participant (I) even expressed concerns about ChatGPT potentially taking too much control over language and knowledge acquisition: “If we use this tool like a slave, it can ruin our learning opportunities, and it could even control our lives and thoughts.”

Overall, the results show that ChatGPT has significant potential as a tool for second language acquisition, provided it is used with structured support. Specific prompts lead to more effective feedback, but learners require guidance to formulate them, particularly those with limited metalinguistic awareness. Cognitive load is an important consideration, as overly complex prompts can overwhelm some learners. While ChatGPT’s usability and feedback are generally appreciated, its full potential is only realized when learners are equipped to interact with the tool effectively. These findings underscore the importance of combining generative AI tools with explicit instructional scaffolding to maximize their educational impact.

Discussion

This study highlights both the opportunities and challenges of integrating generative AI tools like ChatGPT into second language learning for foreign-trained educators. The findings, discussed in relation to the three research questions, offer important implications for practice and future research.

Effectiveness and Precision of Interactions

The results emphasize the critical importance of prompt specificity in eliciting meaningful feedback from ChatGPT. Structured prompts provided in guided classroom settings consistently led to relevant and detailed responses, while general prompts often produced vague corrections that lacked educational value. These findings align with Interaction Theory (Gass & Mackey, 2014), which underscores the necessity of clear and targeted exchanges for effective language learning. Learners who failed to craft precise prompts struggled to engage fully with ChatGPT’s feedback, thereby limiting its potential to facilitate deeper linguistic understanding.

To address this challenge, explicit instructional support is essential. Teachers can model effective prompts and create opportunities for guided practice to help learners develop this skill. Additionally, improvements in AI design, such as integrating built-in scaffolding mechanisms that may prompt the users to clarify their requests, could further enhance the precision and relevance of ChatGPT’s interactions.

Cognitive Demands

The study reveals that the cognitive load associated with using ChatGPT is significantly

influenced by learners' metalinguistic awareness and computer experience, as self-reported in the questionnaire. While complex prompts yielded higher-quality feedback, they also increased extraneous cognitive load, particularly for learners with limited experience. These findings resonate with Cognitive Load Theory (Sweller et al., 2011), which highlights the importance of managing task complexity to support effective learning.

Reducing cognitive demands requires scaffolding and a gradual progression of tasks. Starting with simple activities and incrementally introducing complexity allows learners to build confidence and competence over time. Personalized support through learner profiles that account for individual linguistic and technological expertise can also enhance the learning experience by tailoring instruction to specific needs.

Usability and Perceived Learning Impact

Participants generally appreciated ChatGPT's accessibility and the immediacy of its feedback. However, many found it challenging to use effectively without prior guidance. This gap between perceived usability and actual learning outcomes suggests that while ChatGPT holds considerable potential, its impact largely depends on the context and manner of its integration into the learning process. These findings align with the Technology Acceptance Model (Liu & Ma, 2024), which emphasizes the importance of perceived ease of use and usefulness for the successful adoption of educational technologies.

Teacher involvement is vital in guiding learners' use of ChatGPT to ensure it complements rather than replaces human instruction. This is especially important for adult learners who may be unfamiliar with AI tools. Integrating ChatGPT into broader instructional frameworks, such as activities designed to foster metalinguistic awareness and critical thinking, can further enhance its effectiveness as a language-learning tool.

In summary, while ChatGPT has the potential to transform second language learning, its success depends on effective integration, thoughtful instructional design, and the provision of targeted support to learners. These considerations are essential for leveraging AI tools to their fullest educational potential.

Conclusion

This study highlights the potential of generative AI tools like ChatGPT to enhance second language acquisition for foreign-trained educators by providing personalized feedback and fostering interactive learning. However, the findings also reveal challenges in leveraging this technology effectively, including the need for precise prompts, the cognitive demands placed on learners, and varying levels of technological literacy.

The results suggest that integrating ChatGPT into language learning requires a dual approach: instructional scaffolding to support learners in formulating effective prompts and adapting the technology to better guide user interactions. Explicit training in metalinguistic

awareness and task-specific prompting can empower learners to engage with ChatGPT more effectively, ensuring the feedback they receive is both meaningful and actionable. Additionally, the development of learner profiles could help educators and AI systems tailor feedback to individual needs, reducing cognitive load and enhancing learning outcomes.

Next Steps

Building on these findings, future research should address several important directions to maximize the potential of generative AI tools like ChatGPT in second language education. These directions include both theoretical and practical considerations to ensure that AI is effectively integrated into diverse learning contexts.

1. Longitudinal Studies: Investigating the sustained effects of using ChatGPT on grammatical accuracy, writing fluency, and overall language proficiency over time will provide insights into its long-term educational impact. This research can help determine whether consistent use of generative AI leads to enduring improvements in language skills.

2. AI-Enhanced Scaffolding: Developing adaptive AI systems that may prompt users to refine their inputs can ensure that feedback is more targeted and precise. Such systems could reduce cognitive demands on learners and make interactions with ChatGPT more effective and educationally meaningful.

3. Diverse Contexts and Learners: Exploring how learners from various linguistic and educational backgrounds engage with ChatGPT is critical for understanding its accessibility and equity. Identifying challenges and opportunities for different groups will help make AI-supported learning more inclusive and impactful.

4. Integrative Approaches: Designing curricula that combine generative AI with traditional teaching methods is essential for fostering a balance between learner autonomy and guided instruction. Such integrative approaches can amplify the strengths of both AI tools and human educators.

By addressing these research directions, the field can advance toward a more comprehensive understanding of how generative AI tools can transform language education into a more effective, inclusive, and learner-centered practice.

References

Al-khresheh, M. H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence*, *6*, Article 100218. https://doi.org/10.1016/j.caeai.2024.100218

Bialystok, E. (2001). Metalinguistic aspects of bilingual processing. *Annual Review of Applied Linguistics*, *21*, 169–181. https://doi.org/10.1017/S0267190501000101

Bialystok, E. (2012). The impact of bilingualism on language and literacy development. In T. K. Bahtia & W. C. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (pp. 624–648). Blackwell Publishing. https://doi.org/10.1002/9781118332382

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). *Language models are few-shot learners*. ArXiv. https://doi.org/10.48550/arXiv.2005.14165

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293-332. https://doi.org/10.1207/s1532690xci0804_2

Gass, S. M., & Mackey, A. (2014). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 194–220). Routledge.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 6*, Article 100174. https://doi.org/10.1016/j.caeo.2024.100174

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058-1072. https://doi.org/10.3758/s13428-013-0334-1

Lin, C-C., Huang, A. Y. Q., & Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments, 10*, Article 41. https://doi.org/10.1186/s40561-023-00260-y

Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching, 18*(2), 125–138. https://doi.org/10.1080/17501229.2023.2240316

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351-371. https://doi.org/10.1007/BF02213420

Roehr-Brackin, K. (2018). *Metalinguistic awareness and second language acquisition.* Routledge.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

About the Authors

Sofie Johansson is an Associate Professor, Docent, in Swedish as a Second Language, as well as a senior researcher and lecturer specializing in Multilingualism with a focus on subject didactics. She holds a PhD in Computational Linguistics and Natural Language Processing, with a dissertation on machine learning and part-of-speech tagging. Her research interests include Swedish L2/L3 learning, vocabulary acquisition, text comprehension, language learning and teaching in science and social science education, language and literacy in academic settings, and the application of corpus linguistics in research and education. She has published in *Language Resources and Evaluation Journal*, *Springer*, *Multilingual Matters*, *Oslo Studies in Language*, *US-China Foreign Language*, *Utbildning & Demokrati*, and *Nordiska studier i lexikografi*. She teaches at both undergraduate and postgraduate levels in Swedish, Swedish as a Second Language, and Corpus Linguistics. https://www.gu.se/om-universitetet/hitta-person/sofiejohanssonkokkinakis

Lina Larsson is an Associate Professor in Swedish as a Second Language at the Department of Swedish, Multilingualism, and Language Technology at the University of Gothenburg. She holds a Ph.D. in German with a specialization in literary studies. Her primary teaching responsibilities include introductory language courses for internationally trained teachers, nurses, pharmacists, and doctors. She is currently involved in multiple research and development projects focused on language education for adult second-language learners who are complementing their professional qualifications from outside the EU at a Swedish university. Her main research interests lie in academic literacy for second-language learners, with a particular emphasis on academic reading. She also teaches both undergraduate and graduate courses in Swedish as a Second Language and within teacher education, especially in areas related to literature for second-language instruction and literacy development. https://www.gu.se/om-universitetet/hitta-person/linalarsson

Chapter 4

ChatGPT in ESL Writing: L2 Learners' Practices and Perspectives

Inyoung Na, Mahdi Duris, and Volker Hegelheimer
Iowa State University

This study explores how L2 learners interact with ChatGPT during academic writing revisions and how their individual approaches to using ChatGPT feedback support or hinder L2 writing development. Eight international students enrolled in an ESL writing program at a U.S. university participated in four sessions in a 3-week span, using ChatGPT for feedback on their writing. Data collection included ChatGPT-learner conversation logs, screen recordings, and semi-structured interviews. Analysis revealed diverse interaction patterns through examination of learner prompt types, ChatGPT's feedback types, and learner response types. The potential of these interactions for L2 writing development was assessed through participants' modified output—how learners implemented ChatGPT's feedback into their revisions with both global-level (content, organization) and sentence-level corrections. Case narratives of three participants revealed distinct learner types in their interaction with ChatGPT: Learner, Controller, and Adapter. Analysis revealed that the participant who critically evaluated and selectively integrated ChatGPT's feedback (the Adapter type) showed the most beneficial patterns for autonomy and meaningful L2 development, while over-reliance on ChatGPT's generative capabilities (as observed for the Controller type) posed challenges for deeper engagement with the writing process. This study contributes to the emerging study of AI-assisted writing by identifying specific interaction patterns that enhance or inhibit language development, providing insights for tailoring instructional strategies to maximize ChatGPT's potential in L2 writing pedagogy.

Cite as: Na, I., Duris, M., & Hegelheimer, V. (2025). ChatGPT in ESL writing: L2 learners' practices and perspectives. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 55-79). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.04
©2025 Na, Duris, and Hegelheimer. Published under a CC BY license.

Introduction

The increasing integration of artificial intelligence (AI) in academic writing, defined here as writing tasks assigned in higher education contexts, has changed both student experiences and teaching approaches. Early AI applications in language instruction involved automated writing evaluation (AWE) tools, primarily providing surface-level corrections of linguistic features such as grammar and mechanics. These early tools provided limited scope for learners to engage in iterative dialogue, restricting the breadth, depth, and personalization of the feedback (Barrot, 2024). The recent emergence of generative artificial intelligence (GenAI) tools, exemplified by chatbots like OpenAI's ChatGPT (OpenAI, 2023), has fundamentally changed this landscape, offering real-time, personalized, and interactive feedback through extended dialogue (Carlson et al., 2023; Mun, 2024). These innovative tools not only expand opportunities for students to refine their writing skills outside of classrooms but may also address the challenges faced by teachers who may lack the time to support every learner individually. However, potential issues such as plagiarism and the risk of becoming overly dependent on GenAI technology (Roe et al., 2023) have slowed enthusiasm for GenAI among academic writing teachers.

Despite scholarly attention to plagiarism risks and ethical dilemmas surrounding GenAI-generated text (Godwin-Jones et al., 2024; Roe et al., 2023), few studies have addressed how learners engage with chatbot-generated feedback within an iterative writing process and how such engagement influences their language development. Understanding the extent and manner in which learners critically engage with AI feedback can provide valuable insights into how different learners leverage the use of chatbots and how this impacts their language learning and academic writing skills. Furthermore, interactionist theories of SLA have shown that iterative, dialogic feedback provides important opportunities for learners to engage in negotiation of meaning and produce modified output (Gass & Mackey, 2014; Long, 1996), both of which facilitate deeper cognitive processing and meaningful language development (Swain, 1985). Our study addresses the need for a better understanding of learners' GenAI use in academic research by offering insights from case narratives of learners' interactions with ChatGPT-3.5, examining how individual L2 learners use iterative feedback to revise their academic texts. In doing so, the research aims to identify different learner approaches, which we labeled "Learner," "Controller," and "Adapter" types, and uncover how the strategies of each type affect their language development. By identifying diverse learner practices, educators can better recognize the profiles of "AI-writers" in their classrooms and develop individualized pedagogical strategies to address their needs.

Literature Review

Artificial intelligence integration into writing instruction has rapidly evolved in recent years. Early forms of AI-supported writing instruction primarily involved automated writing evaluation (AWE) tools, whose main strength was providing immediate corrective feedback on grammar, mechanics, and linguistic accuracy (Huawei & Aryadoust, 2023). These tools, such as *Grammarly* and ETS' *Criterion*, typically offered surface-level corrections but often overlooked higher-order writing skills, such as argument cohesion and critical analysis (Godwin-Jones, 2022). Moreover, earlier AWE systems were largely product-oriented, focusing primarily on helping learners polish final drafts rather than supporting deeper cognitive engagement

throughout the writing process (Barrot, 2024). As a result, students' opportunities for interactive dialogue with AI were limited, restricting the depth and personalization of the feedback they received (Barrot, 2024).

Despite their limitations, these AWE tools were generally positively perceived by students due to their immediate and targeted linguistic corrections. However, studies raised concerns regarding students' potential overreliance on automated corrections, warning that excessive dependence on these systems could undermine students' writing autonomy and hinder their development of independent revision skills (Zhai & Ma, 2022). Additionally, Roe et al. (2023) highlighted ethical challenges related to plagiarism fostered inadvertently by specific automated tools (e.g., machine translation [MT], digital writing assistants [DWA], and automated paraphrasing tools [APTs]), including translation plagiarism and back-translation practices, further complicating determination of the authenticity and authorship of writing tasks.

In late 2022, the emergence of generative artificial intelligence (GenAI) tools, such as ChatGPT, fundamentally changed the landscape of AI-assisted writing support (Chapelle et al., 2024). Unlike earlier systems, these GenAI tools provide interactive and personalized assistance, enabling learners to engage in extended dialogue with AI to support various stages of their writing process, including brainstorming, structuring arguments, revising content, and editing language (Mun, 2024; Wang, 2024). Researchers have indicated that GenAI tools not only enhance learners' grammatical accuracy and lexical diversity but also help manage cognitive load, facilitate immediate iterative feedback, and promote positive affective experiences toward writing tasks (Mun, 2024; Wang, 2024). Additionally, blended instructional approaches, where instructors' guidance is combined with GenAI, show promise in helping students benefit from iterative development processes while remaining aware of potential inaccuracies in AI output (Escalante et al., 2023; Hutson et al., 2024).

Altogether, these findings underscore both the potential and the complexities of AI-based writing support. When leveraged conscientiously, GenAI chatbots can provide dialogic, adaptive feedback that has been found to foster learner autonomy (Carlson et al., 2023). Yet, the misuse of GenAI has also revealed new ethical and pedagogical challenges—particularly in ensuring originality and fostering genuine skill development (Roe et al., 2023; Wang, 2024). Ideally, further inquiry into how learners with varied needs adopt AI writing tools can help educators develop nuanced guidelines that balance independent writing with academic integrity.

To understand how different learner-ChatGPT interaction patterns might contribute to language development, we draw on an interactionist perspective (Gass & Mackey, 2006; Long, 2007), and we view learning as emerging through meaningful interaction between the learner and ChatGPT. Within this framework, we consider modified output—the revisions learners make in their writing as a result of receiving and processing AI feedback—as a potential indicator of language learning. Since ChatGPT is inherently reactive rather than proactive, the learning potential largely depends on how learners prompt and interact with it. This learner-AI co-construction process can vary across individuals, with some learners engaging more deeply with feedback than others and demonstrating different levels of metacognitive awareness (Teng, 2025; Yan & Zhang, 2024). By examining these different interaction patterns, we can better understand which approaches may support autonomous writing development and language learning. The present study explores language learners' interaction patterns through case narratives of learners'

interactions with ChatGPT during their writing process. It seeks to identify underexplored dimensions of learners' use of AI-driven writing support and their implications for teaching and learning practices. This study was guided by the following research questions:

1. How do individual L2 learners interact with ChatGPT during academic writing revisions?
2. In what ways do L2 learners' individual approaches to using ChatGPT feedback support or not support L2 development?

Methods

Context and Participants

The data for this study were drawn from Na et al. (2025), which included eight international students (male = 6, female = 2; Age = 18 –25) from diverse first language backgrounds at a large Midwestern U.S. university. All participants completed background questionnaires online, and their biodata is summarized in Table 4.1. They were recruited from a two-course ESL writing sequence: five students were graduate students and were enrolled in the lower-level course (B), and three undergraduate students were enrolled in the higher-level course (C). Based on self-reported English scores, participants could be divided into three proficiency levels: lower proficiency (B05), intermediate proficiency (B01, B02, C07, and C08), and higher proficiency (B03, B04, and C06). Four participants (B02, B03, B04, and C07) reported prior academic writing experience in English-medium education in their home countries, while the others (B01, B05, C06, and C08) had minimal experience with academic writing in English, coming from countries where English was not the primary medium of education. All participants, except B05 (who reported being inactive with respect to AI use), were using the basic version of ChatGPT (ChatGPT-3.5) on either a weekly or daily basis at the time of data collection. The majority of participants had approximately 7-8 months of ChatGPT experience at the time of data collection, including B03, C06, C07, and C08, who also used it for writing assistance. In contrast, B01 and B05 reported minimal prior use. C08 had the most extensive experience using ChatGPT since its public release. All participants were in their first semester in the U.S. at the time of the data collection.

Table 4.1

Characteristics of Participants

ID	Sex	Age	Student /Course Levels	TOEFL/IELTS/etc. (converted to TOEFL score in parentheses)	Major	Type of ChatGPT user
B01	Male	25	G/B	TOEFL 84	Community and Regional Planning	Weekly

Table 4.1 (continued)

Characteristics of Participants

B02	Female	25	G/B	IELTS 6.5 (79-93)	Information Systems and Business Analytics	Weekly
B03	Male	24	G/B	TOEFL 99	MBA	Daily
B04	Male	23	G/B	IELTS 7 (94-101)	Industrial Design	Daily
B05	Male	29	G/B	Duolingo 105 (70-75)	Mechanical Engineering	Inactive
C06	Female	21	U/C	Duolingo 130 (98-103)	Statistics	Daily
C07	Male	21	U/C	Cambridge-B (80-96)	Mechanical Engineering	Weekly
C08	Male	18	U/C	SAT 590	Computer Engineering	Weekly

Note. B and C indicate whether the participant was from the lower writing course (B) or higher writing course (C); U=Undergraduate and G=Graduate. Direct SAT-to-TOEFL score conversion is not available; however, an SAT score of 590 approximates an intermediate TOEFL score.

Data Collection

Four data collection sessions were conducted individually with participants over three weeks in a private lab. The first session included a tutorial, the first three sessions involved a revision task, and the final session consisted of semi-structured interviews along with stimulated recalls. Each session lasted between 45 and 60 minutes.

Tutorial

In the first session, participants watched a 5-minutes tutorial on how to use ChatGPT for essay revisions. This tutorial covered the following: navigating the ChatGPT interface, inputting text for feedback, using basic prompts for general feedback (e.g., “Can you give me general feedback on this essay?”), employing more specific prompts to target their needs (e.g., “Can you provide feedback on the organization of my essay?” and “Are there any grammatical errors?”), and emphasizing the importance of interactive dialogue with ChatGPT, such as asking follow-up questions for clarifications or examples to enhance feedback. As the purpose of the study was to observe how learners would interact with ChatGPT independently, the tutorial was deliberately kept brief to avoid influencing their natural prompting behavior.

Revision Tasks

Following the tutorial, participants were asked to write a short biographical essay of 150–200 words in 20 minutes. This genre was chosen because it is the first assignment students are asked to complete in the ESL writing courses from which they were recruited. They then revised it using feedback obtained from ChatGPT in another 20 minutes. In the second and third sessions, participants brought different texts written prior to the session, where they were asked to revise them with ChatGPT. They submitted their texts to ChatGPT-3.5, prompted it for feedback, and revised their texts in a Google Docs document. Participants worked through the feedback until they were satisfied with their writing or until the 40-minute time limit was reached. Completed drafts were submitted, and conversation logs between the participants and ChatGPT were automatically archived in the ChatGPT system. All sessions were screen-recorded for use in the stimulated recall interviews conducted during the final session.

Semi-structured Interviews/ Stimulated Recalls

In the final session, participants engaged individually in semi-structured interviews and stimulated recalls. The interviews explored their interactions with ChatGPT, perceived effectiveness, and attitudes toward using it for language learning. Additional questions were asked to contextualize their responses, including the text they had provided, prior experiences with ChatGPT, and their English writing background.

Stimulated recalls used researchers' observations of screen recordings. Participants reviewed selected parts of the recordings and explained the reasons for their actions, such as ignoring feedback or typing ChatGPT's suggestions instead of copying them. Questions also assessed whether participants noticed errors or changes made by ChatGPT and whether they learned from the feedback. All interviews and recalls were audio-recorded and transcribed.

Data Coding

To address RQ1 about how L2 learners interact with ChatGPT during academic writing revisions, we identified three variables as important for the description:

- **Prompt type**: The prompts learners used or the specific interactions they sought with ChatGPT (e.g., asking for feedback, revision request).
- **Feedback type**: The types of feedback or output ChatGPT produced in response to the prompts (e.g., direct correction, suggestion, revision).
- **Response type**: How learners reacted to ChatGPT feedback (e.g., no modified, modified, and full modified output)

Prompt type and feedback type were identified from the student-ChatGPT conversation logs, while response type was coded from the screen recordings. These data were integrated into a single Excel file to examine interaction patterns containing the three moves. The finalized prompt, feedback, and response types, and their descriptions are presented in Tables 1–3 in Appendix A. The "Feedback (asking)" prompt type was further coded to identify the specific kinds of feedback participants sought. These included various language aspects (e.g., mechanics, grammar/sentence correctness, cohesiveness), general feedback (i.e., overall comment on the text

as a whole), specific feedback (i.e., specific comment on a particular portion of the text), and others.

To address RQ2 regarding L2 writing development, interaction patterns for individual students identified from RQ1 were considered. Specifically, we focused on three response types following Gurzynski-Weiss and Baralt (2015) and Ranalli (2021): (1) "Modified output," where participants implemented ChatGPT's feedback with their own modification; (2) "No modified output," where feedback was dismissed; and (3) "Full modified output," where participants accepted ChatGPT's revisions in full, typically through direct copying without making further modifications. This coding framework provided a basis for assessing learning processes. The three types of output responses were further coded to reveal where modifications were made at five linguistic levels—mechanics, grammar, lexical, structure, and content. All coding was conducted by the first and second authors. The first author initially coded the data and trained the second author on the coding scheme. The interrater reliability coefficient was 0.98 for prompt and feedback types and 0.97 for response types, calculated using Cohen's Kappa. Disagreements were resolved through discussion until a full agreement (100%) was reached.

Additionally, to provide a more detailed perspective of the learners' experiences, we used transcripts of the stimulated recalls and semi-structured interviews. The transcripts were coded and analyzed following Corbin and Strauss (2007), which involved open coding (identifying relevant units and initial codes for analysis) and axial coding (making connections between codes to create categories). The coding scheme was developed through discussion among the authors, resulting in a final set of eight codes: effectiveness/ineffectiveness of feedback, disliking/liking about the feedback, learning/no learning, perception of ChatGPT, emotional reactions, trust/partial trust, preference, and continued use. The dataset was then recoded by the first author using this final set of codes.

Data Analysis

To address RQ1 on how L2 learners engage with ChatGPT during academic writing revisions, percentages were calculated for the occurrence of each type of prompt, feedback, and response, as coded in the student-ChatGPT conversation logs and screen recordings for each participant. Additionally, the average viewing time and the number of turns between ChatGPT and the learner were calculated using session data exported from ChatGPT, which included accurate timestamps for the interactions. Viewing time was measured as the elapsed time from when participants began viewing ChatGPT's feedback to when they entered their next prompt. This analysis helped identify patterns and trends of individual engagement with ChatGPT's feedback.

To address RQ2, we used response types as indicators of learning outcomes to evaluate which individual learner approaches supported or did not support L2 development. This analytical framework enabled us to determine which interaction patterns had the potential for meaningful language development versus surface-level editing or mere copying-and-pasting behavior. Evidence of learning was considered strongest for "Modified output," while "Full modified output" and "No modified output" were considered less desirable for learning as they bypassed the cognitive engagement necessary for actual language acquisition (Ranalli, 2021). Qualitative analyses involved writing analytic memos for each participant based on the coded transcriptions and biodata from the interviews. While response categorization provided quantifiable evidence of learning behaviors, the interview data offered deeper insights into

learners' metacognitive awareness and self-reported perceptions of development. The quantitative findings from RQ1 and qualitative analyses were combined to select three participants, B03, B05, and C08, as foci for case-study narratives, as they were "very informative cases that represent a phenomenon of interest intensively" (Teddlie & Tashakkori, 2008, p. 175).

Results and Discussion

Nature of Interaction Patterns between Student and ChatGPT (RQ1)

The analysis addressing RQ1 examined how L2 learners interacted with ChatGPT. Tables 2, 4, and 5 present the distribution of prompt types (Table 4.2), feedback types (Table 4.4), and response types (Table 4.5) across three revision sessions for each participant. As shown in Table 4.2, "Feedback (asking)" (e.g., C06: *"Can you give me a feedback about this essay?"*; C08: *"Where can I add counterarguments in the essay?"*) was the most frequently used prompt type (142; 45.25%), which was expected given that participants were instructed to prompt ChatGPT for feedback during revisions. Table 4.3 further breaks down the feedback (asking) type into subcategories, highlighting the specific language aspects participants addressed with ChatGPT. "Revision request" (e.g., B01: *"Could you rewrite the revised text with simpler language?"*; B03: "*Rewrite the fourth paragraph, incorporating your suggestions and feedback. Try to make it concise."*) was the second most common (75; 23.73%), reflecting participants' reliance on ChatGPT to rewrite their text. "Follow-up" prompts (e.g., B01: *"Show me exactly what you did and in which sentences"*; C08: "*Can you provide me some examples?"*) were the third most frequent (65; 23.10%), highlighting iterative engagement, where participants re-queried ChatGPT to refine output (Mun, 2024; Wang, 2024). B03 stands out for having made the highest number of prompts in total (61), nearly half of which were revision requests (27). In contrast, C08 requested feedback the most frequently (33) among all participants but made the fewest revision requests (2). Interestingly, 14 of C08's "Feedback (asking)" prompts resulted in "Revision" feedback, even though revisions were not explicitly requested. In total, C08 received revised text from ChatGPT 16 times, as shown in Table 4.4.

Table 4.2

Frequency of Prompt Types by Participant (n=8)

Prompt	B01	B02	B03	B04	B05	C06	C07	C08	Total
Feedback (asking)	26	12	15	7	17	23	9	33	142
Revision	7	11	27	13	4	5	6	2	75
Follow-up	10	7	19	2	4	6	4	13	65
Search	3	1	0	0	1	5	2	0	11
Meaning negotiation	2	0	0	1	3	3	0	0	9

Table 4.2 (continued)

Frequency of Prompt Types by Participant (n=8)

Evaluation	2	5	0	0	0	0	0	0	7
Generation	0	4	0	2	0	0	0	0	6
Total	50	40	61	25	28	42	21	48	315

Note. Follow-up was double-coded with any other type.

Table 4.3

Breakdown of 'Feedback (asking)' Types (n=8)

Feedback (asking)	B01	B02	B03	B04	B05	C06	C07	C08	Total
general	9	5	6	2	2	3	2	3	32
specific	1	2	6	2	2	8		7	28
grammar/ sentence	7	1			3	3	3		17
vocabulary	6		2		3			3	14
content		1			2	3	2	3	11
structure	1	1	1	1	2	2	1	1	10
delivery					1			8	9
cohesiveness	1			1	2	1		1	6
mechanics		2				1			3
others	1			1		4	1	7	13
Total	26	12	15	7	17	25	9	33	143

Many "Feedback (asking)" prompts led to ChatGPT providing fully revised text. Therefore, among the nine feedback methods used by ChatGPT, "Revision" was the most frequently used method (114; 40.43%), followed by "Suggestion" (75; 26.60%) and "Summative comment" (44; 15.10%) (Table 4.4). While the revision type demonstrates the adaptability and utility of an LLM in tailoring its responses to specific revision needs, it also highlights a potential risk of learner over-reliance. This reliance may limit active language learning and raise

concerns about academic integrity (Roe et al., 2023), as learners may bypass the revision process entirely by simply copying and pasting ChatGPT's output, or what is termed as "full modified output."

Table 4.4

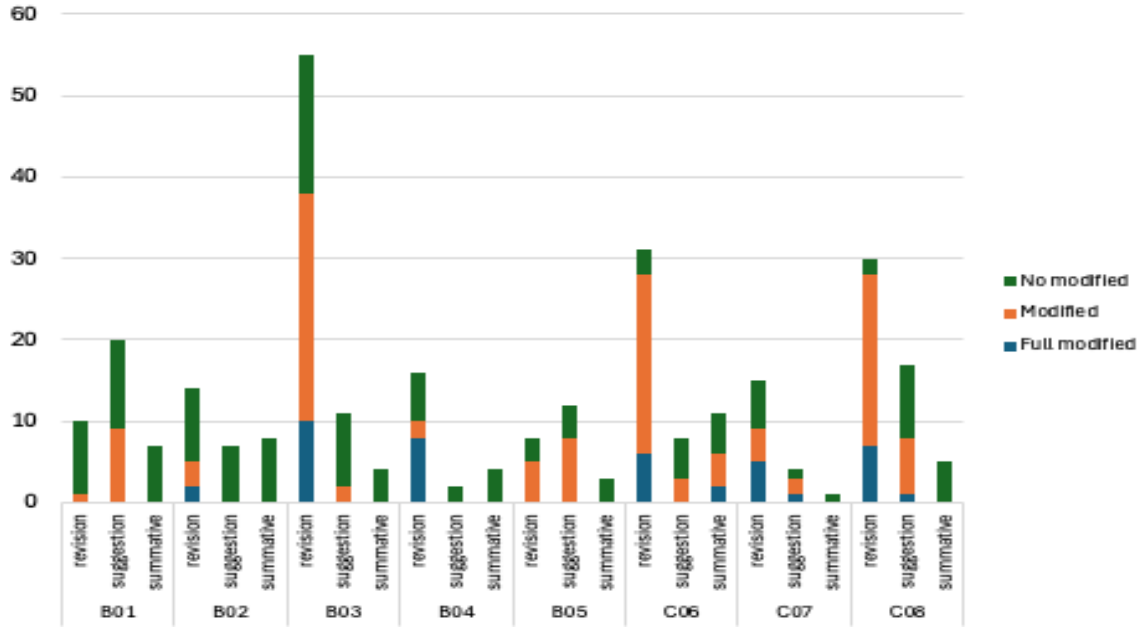
Frequency of Feedback Methods Given by ChatGPT by Participant

Feedback	B01	B02	B03	B04	B05	C06	C07	C08	Total
Revision	10	12	31	14	6	16	10	16	114
Suggestion	16	7	12	4	11	7	2	16	75
Summative	7	8	4	6	3	10	1	5	44
Direct correction	4	3	0	0	2	2	1	1	13
Information	3	1	0	0	1	6	1	0	12
Metalinguistic	1	0	0	0	6	0	0	5	12
Summary	0	1	0	4	0	0	0	1	6
Generation	0	4	0	0	1	0	0	0	5
Locating	0	0	0	0	0	0	1	0	1
Total	41	36	47	28	31	41	16	44	282

Figure 4.1, which shows the distribution of response types (no modified, modified, and full modified output) across the three most frequently used feedback methods, illustrates this pattern. The full modified response type is more prevalent following "Revision" feedback but seldom occurs following "Suggestion" feedback (e.g., ChatGPT providing a bulleted list of improvement points such as *"Vocabulary enhancement: introduce more diverse and sophisticated vocabulary where appropriate to add depth and nuance to the text"*) and "Summative Comment" feedback (e.g., ChatGPT providing summarized evaluative comments, both positive and/or negative, such as *"The biography is clear and well-structured"*) feedback types.

Figure 4.1

Distribution of Response Types across the Top Three Categories of Feedback Received from ChatGPT



Suggestions, and particularly summative comments, mostly resulted in no modified output, which was recorded as the most frequent among the three response types (Table 4.5). This may imply that participants use suggestions and summative comment types of feedback more as a reference rather than a direct source for rewriting. Alternatively, it could reflect that these feedback categories were perceived as less useful or actionable, yet further examination of participants' perceptions is needed. For B03, while modified output dominated, full modified responses were still present in the "Revision" category, suggesting occasional reliance on ChatGPT's output. In contrast, for B05, this feedback did not lead to any full modified responses. Across all sessions, B05 recorded only one instance of full modified output—the lowest among all participants, as shown in Table 4.5.

As shown in Table 4.5, the proportion of no modified output is the highest, indicating a possible perception of inapplicability or irrelevance, as later confirmed in interviews. In contrast, the proportions of modified and full modified responses varied noticeably between individuals, as visually represented in Figure 4.2 below, which shows the distribution of response types across participants. For example, B03 and C07 exhibited a much higher ratio of full modified responses, indicating a greater tendency to adopt ChatGPT's revisions verbatim, while B05 favored almost entirely no modified or modified responses. This variation underscores that, although all participants engaged with ChatGPT's feedback, their reliance on AI-generated revisions differed considerably, reflecting individual revision strategies and levels of reliance on AI assistance.

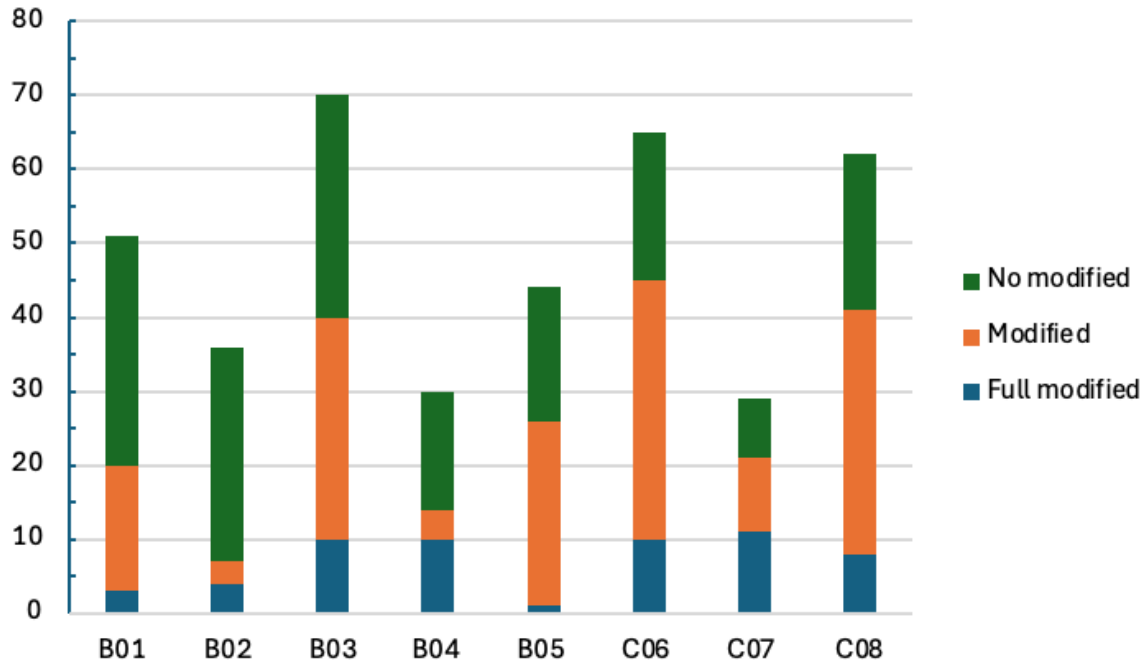
Table 4.5

Frequency of Response Types by Participant

Response	B01	B02	B03	B04	B05	C06	C07	C08	Total
Full modified	3	4	10	10	1	10	11	8	57
Modified	17	3	30	4	25	35	10	33	157
No modified	31	29	30	16	18	20	8	21	173
Total	51	36	70	30	44	65	29	62	387

Figure 4.2

Distribution of Response Types across Participants



Time spent engaging with ChatGPT's feedback (Table 4.6) showed a maximum of 240.52 seconds on average per feedback for B05, who had 8 turns with ChatGPT per session, compared to B03, who made the most turns (16) with a lower average of 171.24 seconds per feedback. Similarly, C08 exhibited a relatively efficient engagement style, with an average viewing time of 147.90 seconds and 13.7 turns. However, C08 demonstrated a distinct approach by making only two explicit revision requests as opposed to B03's frequent reliance on revisions, making 27 revision requests (Table 4.2). In contrast, B05 represents a less efficient engagement

style, with the longest average viewing time but the fewest instances of full modified output among all participants. These observations informed the selection of B03, B05, and C08 as the focus of case-study narratives to answer RQ2, allowing for a comparative exploration of their diverse interaction strategies and prompting behaviors.

Table 4.6

Average Viewing Time per Feedback and Average Number of Turns across Sessions

Person	Viewing Time (in seconds)	Number of Turns
B01	97.60	15.0
B02	139.16	12.3
B03	171.24	16.0
B04	121.80	9.7
B05	240.52	8.0
C06	167.43	12.0
C07	230.07	5.7
C08	147.90	13.7

Case-study Narratives: Potential for Learning (RQ2)

For readability, pseudonyms are introduced for the selected participants in this section: B03 (Alex), B05 (Sam), and C08 (Ryan). Participant identifiers (e.g., B03) are maintained for cross-referencing with the tables presented in RQ1.

Case 1: "Controller" (Alex)

Alex, a 24-year-old from India, was a master's student in industrial design. He had been placed in a lower-level ESL writing course by the university's English placement test and had spent three months in the program. English had been his primary language of instruction in India. When asked how he felt about academic writing in English, Alex noted that he often struggled with organizing his ideas and writing longer essays, though he did not feel underconfident about his writing abilities. He described ChatGPT as "a part of his thing" as a frequent ChatGPT user since he started using it eight months ago, having become familiar with its functionalities and primarily using it for writing assistance.

As noted, Alex made the highest number of turns and prompts among all participants, indicating his high engagement with ChatGPT (B03 in Tables 2 and 6). Despite his numerous

prompts, nearly half were revision requests and follow-ups, reflecting his strategic approach to iterative refinement until he was satisfied with ChatGPT's output. This process often resulted in copying and pasting fully revised drafts generated by ChatGPT. Of the 30 instances of modified output Alex produced using ChatGPT's feedback (Table 4.5), the modifications were made at the levels of mechanics (1), grammar (7), lexical (9), structure (6), and content (7).

Alex typically began with prompts like "Rewrite this essay effectively" and iteratively refined ChatGPT's output through multiple follow-ups, such as "Humanize this," "In 200-250 words," "Keep it simple and semi-formal," and "Don't overdo it or exaggerate it." Explaining his approach, Alex said:

> So I just wanted to be as specific as possible with my prompt so that I'll get the clear and correct version out of it [...] I wanted to make sure that what I had in mind would be reflected in ChatGPT. (Alex)

For "Humanize this," Alex highlighted his awareness of ChatGPT's tendency to fabricate and exaggerate sentences, pointing out issues with hallucinations, saying:

> It is giving me some made-up writings and all, made-up sentences and all that we cannot find on the internet or so. So that's why I wanted to say, 'humanize this,' and make it more, make it seem human rather than like something inhumane or something made up. (Alex)

Therefore, by using prompts like "humanize this," Alex believed that ChatGPT provided more accurate and credible information, avoiding the risk of generating "made-up" sentences. Still, he acknowledged, "I have to go through them and edit."

In fact, rather than directly revising his drafts on his own, Alex often asked ChatGPT to rewrite his text and then sought feedback on the AI-generated draft. He further refined the output by replacing exaggerated or overly formal phrases with something that "he" would use, not an AI. Nonetheless, Alex noted how ChatGPT helped him learn new words, identify areas for improvement, structure, and refine his writing, expressing a clear willingness to continue to use ChatGPT for his future writing processes. Alex exemplifies the "Controller" type through his strategic use of iterative prompting to direct ChatGPT's output, often bypassing his own revision process.

Case 2: "Learner" (Sam)

Sam was a 29-year-old from Brazil pursuing a Ph.D. in Mechanical Engineering. He described academic writing in English as a struggle, citing his insecurities with grammar and vocabulary, having begun formal English writing instruction only three months ago in a low-level ESL writing class. Before coming to the U.S., Sam had self-studied English using Duolingo but had no experience with academic writing or standardized exams like TOEFL or IELTS. Describing himself as a novice ChatGPT user, he recalled his failure to get good answers from ChatGPT when attempting to search for topics. He was unfamiliar with prompting and using ChatGPT for writing assistance.

Sam's unfamiliarity with ChatGPT became evident in the first session when he cautiously began the conversation by typing, "Hi, can you please give me some feedback on my biography?" He later noted that the feedback was often too lengthy and not to the point, such as offering only positive comments when he wanted to address weaknesses. By the final session, however, he explained that he had learned to use more specific prompts, such as "Can you please point out the weaknesses in this text?" to receive the targeted feedback from the beginning.

As noted, despite making fewer turns and prompts, Sam spent the most time of any participant engaging with ChatGPT (B05 in Table 4.6). He carefully read the feedback and implemented it on his own, seeking clarifications from ChatGPT when needed. Of the 17 feedback requests, Sam asked for feedback on diverse aspects, ranging from general feedback (2) and specific feedback (2), to more specifically on language aspects, including local concerns such as grammar (3) and vocabulary (3), and global concerns, such as structure (2), content (2), delivery (1), and cohesiveness (2) (Table 4.3). Sam also sought metalinguistic information six times from ChatGPT (Table 4.2). For instance, he asked about how to craft a good "hook" and what "intended audience" meant for academic writing, reflecting his effort to improve his understanding of writing principles through interactions with ChatGPT.

Among the 25 instances of modified output Sam produced (Table 4.5), most were at the content level (13), followed by lexical (6), grammar (3), structure (2), and mechanics (1). Notably, Sam produced the fewest full modified outputs —only once—among all participants (Table 4.5). He mentioned his strong commitment to avoiding plagiarism, which was why he did not copy and paste but paraphrased instead. Another reason was because of his strong learning motivation. When asked, "Why did you not copy and paste?" he said, "Because I wanted to improve this, like paraphrasing skills. Because it's very important in the research, right?"

In one notable instance, ChatGPT inaccurately identified the use of a singular verb in Sam's sentence, "Although learning a new language demands..." as incorrect, suggesting it be replaced with "demand." Sam initially thought "demands" was correct but followed ChatGPT's suggestion anyway. Later, when the researcher confirmed that the original sentence was correct, Sam reflected on how ChatGPT had caused confusion despite his doubts. Despite this and ChatGPT's tendency to provide overly general feedback, Sam said he would continue to use ChatGPT, especially when he needed help brainstorming for ideas or overcoming writing difficulties, stating, "[...] It helps…Any help? It's welcome." Sam represents the "Learner" type by approaching ChatGPT as a learning tool, carefully processing feedback and implementing suggestions manually to develop his own skills.

Case 3: "Adapter" (Ryan)

Ryan, an 18-year-old from Nepal, was a freshman undergraduate studying Computer Engineering, placed in the higher-level ESL course. He noted that academic writing in English was a "new thing" for him, as he had not been taught how to write essays in a structured way. An adept ChatGPT user, Ryan began using ChatGPT soon after it was made public, primarily to address his weaknesses in grammar and enhance the flow and tone of his writing—goals and focus he also brought to the study sessions.

Ryan made the highest number of feedback requests among participants, but only two revision requests (C08 in Table 4.2). His 33 feedback requests spanned a diverse range of

aspects, including general feedback (3), specific feedback (7), vocabulary (3), content (3), structure (1), delivery (8), cohesiveness (1), and others (7) (Table 4.3). He typically began his prompts broadly to identify general issues before refining them through iterations, focusing on specific aspects or sections of his writing. Notably, towards the end of his sessions, he posed highly precise questions such as "What is a better ending for the following paragraph?" and "Where can I add counterarguments in the essay?"—reflecting his desire for actionable and detailed guidance.

Of the 33 instances of modified output Ryan produced (Table 4.5), the modifications were made across mechanics (2), grammar (6), lexical (11), structure (7), and content (7). 21 of them were based on 'Revision' feedback, even though he had only asked for feedback rather than fully revised text. Expressing his annoyance, he stated, "So I wanted it to give me an idea of how to revise it; it actually revised the essay for me. So that was the issue in here." However, he attributed this outcome to his own prompting, adding, "But it was my fault because I didn't give it the proper prompt." Instead, he noted that he sought to learn from ChatGPT's revisions, particularly when and where to use certain words in sentences.

Notably, Ryan demonstrated a balanced integration of ChatGPT's suggestions while maintaining his tone and authorship. He explained, "I didn't want to change everything… I try to keep it between my tone and ChatGPT's tone." His strong sense of ownership over his writing was evident when he remarked,

> I know some of the things that I write are working. When I put things in ChatGPT, it tries to take it away from the thing I'm trying to write, posting a bunch of stuff from the internet, which I don't think is going to fit in there. (Ryan)

While Ryan expressed that he would continue to use ChatGPT, he emphasized a cautious approach: "So I won't totally rely on that for my writing tasks… But yeah, I'm still going to use it, but not totally, like, maybe like 20%." Ryan embodies the "Adapter" type by selectively integrating ChatGPT's suggestions while maintaining his own voice and critical awareness of the AI's limitations.

Learner Types in ChatGPT Interaction

The case narratives of Controller (Alex), Learner (Sam), and Adapter (Ryan) illustrate varying interaction types with ChatGPT that demonstrate different degrees of engagement aligned with interactionist theories of SLA (Gass & Mackey, 2014; Long, 1996). Sam's "Learner type" reflects cautious exploration for gaining technological competence (Yan, 2024). Sam initially used general prompts—expressing frustration with overwhelming feedback—but later developed more focused prompts that better addressed his needs. With a strong motivation to improve his English writing, he carefully read ChatGPT's feedback and implemented it in his writing on his own. He avoided copying-and-pasting, producing only 1 full modified output (the least among all participants). However, he often accepted ChatGPT's feedback uncritically, as seen when he even followed an incorrect suggestion, which may be related to his reported low English proficiency and/or limited prior experience with ChatGPT, as indicated in his background information.

Alex's "Controller type" leverages strategic prompting techniques for efficiency, though often at the expense of deeper engagement with feedback. His frequent use of full modified output highlights the potential for over-reliance on AI tools and the associated risks of plagiarism. This would indicate surface-level interaction and limited cognitive depth associated with excessive reliance on AI (Roe et al., 2023).

Ryan's "Adapter type" shows a clear awareness of both his own writing strengths and ChatGPT's limitations, demonstrating a level of metacognition that likely enhanced his learning outcomes (Teng, 2025). Ryan selectively adopted ChatGPT's suggestions while maintaining his personal writing goals. He produced substantially modified output by revising his own text with ChatGPT's feedback, which indicates a deeper cognitive processing of feedback rather than mere acceptance. Among these, the Adapter approach appears most conducive to meaningful L2 writing growth, revealing how learners can maintain ownership of their writing while benefiting from generative AI's strengths (Carlson et al., 2023). These distinctions underline that the potential of AI tools like ChatGPT for language learning is maximized when learners actively negotiate meaning and engage critically with iterative feedback, confirming the theoretical expectations discussed in our introduction.

Our findings have important pedagogical implications for instructors integrating ChatGPT into L2 writing classrooms. As teachers and researchers have found in the past, learners need guidance in their use of technology for learning (Hubbard, 2013). In this case, instructors should provide a comprehensive "Writing with AI" orientation that goes beyond the basic technical functionality of ChatGPT, including effective prompting strategies, responsible AI use, and critical evaluation of AI-generated feedback in support of L2 writing development. Our research highlights the need for scaffolded integration rather than unsupervised use before gradually giving students more autonomy as they demonstrate responsible AI use patterns. For Controller-type learners who tend toward over-reliance, specific guardrails are needed, such as requiring students to document how they have processed AI feedback. Additionally, assessment practices should value process over product, awarding points for thoughtful engagement with feedback rather than just final text quality. These approaches aim to foster Adapter-type behaviors across all learners—the pattern our research identified as most conducive to language development—by promoting autonomous but critical AI use. This should help students maximize benefits while mitigating risks of over-reliance or passive acceptance of AI-generated content.

Conclusion

The goal of this study was to obtain detailed qualitative insights into how L2 English learners interact with ChatGPT for writing assistance. We analyzed ChatGPT-learner conversation logs, screen-capture recordings, and interviews across three revision sessions. Interaction patterns focused on quantifying types of learner prompts (e.g., asking for feedback, requesting revision), the nature of feedback received (e.g., suggestions for improvement, summative comments, direct corrections), and learner responses to the feedback (i.e., no modified, modified, and full modified output). Through case narratives grounded in quantitative data and interviews, we identified three distinct learner types with respect to their ChatGPT interaction: Learner, Controller, and Adapter. These types encapsulate the different tendencies for strategies guiding learners' interactions with ChatGPT for feedback.

This study provides key insights into the importance of how learners interact with ChatGPT and similar AI tools rather than merely focusing on whether they utilize these technologies. While ChatGPT offers valuable support for L2 writers, its benefits appear to be maximized when learners embrace an Adapter approach—critically evaluating and selectively implementing feedback—rather than relying on it as a shortcut. An Adapter approach fosters both autonomy and advanced revision strategies. However, the study also highlights notable limitations**,** particularly when learners become overly dependent on AI-generated text (Controller type), which may hinder deeper engagement with the writing process. These findings underscore that AI tools like ChatGPT are neither inherently beneficial nor detrimental to L2 development; rather, their impact depends largely on how they are used and the instructional scaffolding provided. Overall, this study contributes to AI-assisted L2 writing research by identifying specific interaction patterns that either enhance or inhibit language development, offering insights for tailoring instructional strategies to maximize ChatGPT's potential while mitigating its risks in language learning contexts. Based on our findings, we propose the following instructional strategies:

- **Explicit Prompting Instruction:** To support *Learner-type students*, instructors should teach them how to craft specific, purposeful prompts. This would help students shift from asking for general feedback to seeking focused input, resulting in more actionable and targeted feedback.

- **Ethical AI Use Training:** For *Controller-type learners*, instructors should address over-reliance on AI and the risk of plagiarism by teaching academic integrity. This can include explicit instruction on proper paraphrasing, citing AI-generated content when used, and distinguishing between acceptable use and unethical copying. Encouraging learners to request more "Suggestion"-type feedback (e.g., "*Can you suggest ways to clarify this argument?*" and "*How can I improve the organization of this paragraph?*") rather than direct revisions (e.g., *"Can you revise this draft?"*) can promote deeper engagement with feedback and reduce over-reliance on AI-generated revisions.

- **Reflective Revision Practices:** To help students develop selective engagement strategies akin to the *Adapter type*, instructors can implement reflective activities where students explain how they used or rejected ChatGPT feedback. This would encourage thoughtful integration of feedback and foster learner autonomy by helping them balance AI assistance with their own voice.

We hope readers gain an understanding of the diverse strategies learners engage with AI tools like ChatGPT and recognize the importance of guiding learners toward adaptive, reflective use for more meaningful L2 writing development. Though our findings identified three distinct learner types—Learner, Controller, and Adapter—it remains unclear whether these represent developmental stages or stable learner preferences. While our data suggested certain learner types, we acknowledge that this classification is preliminary and would benefit from further empirical validation across different contexts and learner populations. As an initial exploration, this study is limited by its short duration, which may not have been sufficient to observe changes in learners' interaction patterns over time. Future research should include longitudinal studies to address this gap and investigate whether these learner types represent developmental stages through which students progress as they gain experience with AI writing tools. Additionally, we

did not explicitly focus on the impact of language proficiency on learners' use of ChatGPT. For example, higher English proficiency may enable learners to formulate nuanced prompts and obtain more tailored feedback. Future research should consider language proficiency as a variable, enabling comparisons between lower- and higher-level proficiency to further enrich our understanding of ChatGPT's role in L2 writing development. Lastly, future studies should explore how instructors can guide learners during AI-supported revisions—for instance, through explicit prompting instruction, reflective revision practices, and training in ethical AI use. Such research can examine how explicit instructor guidance in AI use affects growth in learner autonomy, metacognition, and L2 writing development.

Our findings also highlight important ethical and pedagogical considerations for AI integration in L2 education. As ChatGPT and similar tools become increasingly prevalent in writing classrooms, institutions must develop clear guidelines that balance the benefits of AI assistance with concerns about academic integrity. The different learner types we identified—particularly the Controller approach with its heavy reliance on AI-generated text—raise questions about authorship and authentic language development. Educators should consider implementing process-oriented approaches that integrate AI at different stages of L2 writing—from idea generation and outlining to obtaining feedback and refining drafts—rather than focusing solely on final products. Additionally, explicit instruction in responsible AI use should become part of L2 writing pedagogy, helping students understand when and how to appropriately leverage these tools while developing their own language skills. Institutions might consider developing policies that acknowledge AI as a writing resource while establishing boundaries that preserve meaningful language learning opportunities and academic integrity.

References

Barrot, J. S. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies, 29*(6), 7155–7179. https://doi.org/10.1007/s10639-023-12083-y

Carlson, M., Pack, A., & Escalante, J. (2023). Utilizing OpenAI's GPT-4 for written feedback. *TESOL Journal*, *15*(2), e759. https://doi.org/10.1002/tesj.759

Chapelle, C. A., Beckett, G. H., & Ranalli, J. (2024). Paths for exploring AI in Applied Linguistics. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 1–8). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.01

Corbin, J., & Strauss, A. (2007). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd edition). Sage. https://doi.org/10.4135/9781452230153

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), Article 57. https://doi.org/10.1186/s41239-023-00425-2

Gass, S., & Mackey, A. (2006). Input, interaction and output: An overview. In K. Bardovi-Harlig & Z. Dörnyei (Eds.), *Themes in SLA research* (pp. 3–17). John Benjamins. https://doi.org/10.1075/aila.19.03gas

Gass, S. M., & Mackey, A. (2014). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 180–206). Routledge. https://doi.org/10.4324/9780203628942

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 26*(2), 5–24. http://doi.org/10125/73474

Godwin-Jones, R., O'Neill, E., & Ranalli, J. (2024). Integrating AI tools into instructed second language acquisition. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 9–23). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.02

Gurzynski-Weiss, L., & Baralt, M. (2015). Does type of modified output correspond to learner noticing of feedback? A closer look in face-to-face and computer-mediated task-based interaction. *Applied Psycholinguistics, 36*(6), 1393–1420. https://doi.org/10.1017/S0142716414000320

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies, 28*(1), 771–795. https://doi.org/10.1007/s10639-022-11260-9

Hubbard, P. (2013). Making a case for learner training in technology enhanced language learning environments. *CALICO Journal*, *30*(2), 163–178. https://doi.org/10.11139/cj.30.2.163-178

Hutson, J., Plate, D., & Berry, K. (2024). Embracing AI in English composition: Insights and innovations in hybrid pedagogical practices. *International Journal of Changes in Education, 1*(1), 19-31. https://doi.org/10.47852/bonviewIJCE42022290

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

Long, M. H. (2007). *Problems in SLA*. Lawrence Erlbaum.

Mun, C. Y. (2024). EFL learners' English writing feedback and their perception of using ChatGPT. *Journal of English Teaching through Movies and Media, 25*(2), 26–39. https://doi.org/10.16875/stem.2024.25.2.26

Na, I., Duris, M., & Hegelheimer, V. (2025). Feedback and automated writing evaluation. In Y. Wang, A. Alm, & G. Dizon (Eds.), *Insights into AI and language teaching and learning*. Castledown (pp. 290-313). https://doi.org/10.29140/9781763711600

OpenAI. (2023). *ChatGPT* (Mar 14 Version) [Large language model]. https://chat.openai.com/chat

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, Article 100816. https://doi.org/10.1016/j.jslw.2021.100816

Roe, J., Renandya, W. A., & Jacobs, G. M. (2023). A review of AI-powered writing tools and their implications for academic integrity in the language classroom. *Journal of English and Applied Linguistics, 2*(1), Article 3. https://doi.org/10.59588/2961-3094.1035

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass and C. G. Madden (Eds), *Input in second language acquisition* (pp. 235-253). Newbury House.

Teddlie, C., & Tashakkori, A. (2008). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage.

Teng, M. F. (2025). Metacognitive awareness and EFL learners' perceptions and experiences in utilising ChatGPT for writing feedback. *European Journal of Education, 60*(1), 1–21. https://doi.org/10.1111/ejed.12811

Wang, C. (2024). Exploring students' generative AI-assisted writing processes: Perceptions and experiences from native and nonnative English speakers. *Technology, Knowledge and Learning*, 1–22. https://doi.org/10.1007/s10758-024-09744-3

Yan, D. (2024). Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology, 28*(1), 1–19. https://hdl.handle.net/10125/73597

Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications*, *11*(1), Article 1086. https://doi.org/10.1057/s41599-024-03543-y

Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning, 35*(9), 2817–2842. https://doi.org/10.1080/09588221.2021.1897019

Appendix A

Table A1

Prompt Types and Descriptions

Prompt Type	Description
Asking for Feedback	The participant asks for feedback on their text. *(e.g., provide specific feedback on the fourth paragraph; can you provide me some examples?)*
Revision Request	The participant asks ChatGPT to rewrite the text. *(e.g., make it better and polished in 200-250 words; rewrite the second paragraph with all your suggestions)*
*Follow-up	The participant follows up on ChatGPT's response with a further request or meaning negotiation. *(e.g., I don't feel that giving examples is a good point; with more simple language)*
Information Search	The participant searches for general information not related to the target language. *(e.g., how much CO2 is produced to make one car)*
Evaluation Request	The participant asks ChatGPT to grade their paper. *(e.g., Would you rate this out of 100?)*
Text Generation	The participant asks ChatGPT to generate new text without basing it on their own text. *(e.g., write your response on the findings)*

Note. Follow-up was double-coded with any other type.

Table A2

Feedback Types and Descriptions

Feedback Type	Description
Revision	ChatGPT produces the revised text.
Suggestion	ChatGPT gives suggestions for improvement, usually providing a bulleted list of points. *(e.g., Vocabulary Enhancement: Introduce more diverse and sophisticated vocabulary where appropriate to add depth and nuance to the text.)*
Summative comment	ChatGPT provides a summary of evaluative comments on the writing. *(e.g., The biography is clear and well-structured.)*
Direct correction	ChatGPT indicates that there was an error in the writing with the correct form. *(e.g., Changed "comes" to "stems" for better tense agreement.)*
Information	ChatGPT provides general knowledge not related to the target language. *(e.g., Red symbolizes passion, energy, and good fortune in Korea.)*
Metalinguistic information	ChatGPT gives information on language and writing. *(e.g., Subject-Verb Agreement: Ensure that the subject and verb agree in number and tense; Question Hook: Pose a thought-provoking question that encourages readers to consider the impact of technology on children's development.)*
Summary	ChatGPT provides a summary of the content of the participant's writing. This usually is a breakdown of the text. *(e.g., Here's the breakdown of the essay: ...)*
Generation	ChatGPT generates the text upon request, not based on the original text. *(e.g., Certainly, I'll provide responses to two of the findings from the study: ...)*
Locating	ChatGPT visually highlights the location of what it's indicating with bolded text. *(e.g., Here's the highlighted thesis statement: ...)*

Table A3

Response Types and Descriptions

Response Type	Description
No modified output	The learner does not make any modifications to the text or follow up with another request (double-coded with the prompt type, "Follow-up"
Modified output	The learner partially implemented ChatGPT's feedback at five levels: - Grammar (e.g., verb tense/form, SV agreement, plurals) - Mechanics (e.g., citation format, punctuation) - Lexical (e.g., replacement of words or phrases) - Structure (e.g., sentence and paragraph structures) - Content (e.g., additions, deletions, modified ideas)
Full modified output	The learner fully implemented ChatGPT's feedback (i.e., copy and paste)

About the Authors

Inyoung Na (inyoungn@iastate.edu) is a Ph.D. student in Applied Linguistics & Technology at Iowa State University. Her research focuses on the use of technology in second language speaking, second language writing, and language testing and assessment, with particular emphasis on generative AI-based spoken dialogue systems and Automated Writing Evaluation (AWE). https://engl.iastate.edu/profiles/inyoung-na/

Mahdi Duris (mduris@iastate.edu) holds a Ph.D. in Applied Linguistics & Technology at Iowa State University, focusing on the use of technology in second language pronunciation, second language writing, and AI applications in education. His research interests center on pronunciation, phonetics, and phonology, especially regarding the automation of intelligibility ratings, as well as Automated Writing Evaluation (AWE) and the broader integration of AI in educational contexts. https://mahdiduris.com.

Volker Hegelheimer (volkerh@iastate.edu) is a professor of applied linguistics in the Department of English at ISU. He teaches courses on technology in language teaching and research, language assessment, and research methodology. His research interests include applications of technology in language learning and language testing. https://engl.iastate.edu/profiles/volker-hegelheimer/

Chapter 5

Investigating the Influence of ChatGPT Writing Guidance on Peer Review Assessments: An Exploratory Study

Danilo Calle
Iowa State University

This study explored the impact of ChatGPT as an Automated Writing Evaluation (AWE) tool on ESL students' peer feedback language and criteria during an online writing course. Eight students engaged in a feedback loop across three writing tasks, receiving ChatGPT-generated feedback before reviewing peers' work. Data were compiled into two corpora, analyzing students' and ChatGPT's linguistic complexity, lexical diversity, and evaluation focus. The results revealed significant linguistic differences: ChatGPT feedback exhibited higher sentence complexity and lexical diversity, while student feedback emphasized epistemic markers and lexical repetition. Interaction with ChatGPT influenced students' language, leading to increased nominalizations and a more informational style over time. ChatGPT has become a frequent feedback tool in the writing process, but instructors do not often consider the effect of recurrent interaction with AI on students' language and considerations of appropriate feedback. This exploratory study addresses the current gap in understanding the influence of ChatGPT on students' feedback considerations.

Cite as: Calle, D. (2025). Investigating the influence of ChatGPT writing guidance on peer review assessments: An exploratory study. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 80-102). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.05
©2025 Calle. Published under a CC BY license.

Introduction

Corrective Feedback (CF) is widely used in Second Language (L2) learning to guide students and support them on how to use language effectively in their writing. Although providing personalized feedback to students addresses their individual needs in a way that benefits their improvement, the writing process can be time-consuming and demanding for teachers when the class conditions, resources, and time availability are not convenient. Fortunately, today's view of classrooms as spaces for collaboration and innovation allows for different types of corrective feedback to be implemented with the purpose of promoting writing skills improvement while supporting teachers with their workload. The integration of teacher feedback, peer feedback, and automated feedback has become a means for dynamically mediating the writing process to engage learners with different error correction sources.

Different approaches to providing feedback have been proposed to take into consideration students' characteristics, error types, and classroom contexts (Ellis, 2009). However, there is no universal method for providing feedback that works for all students in all circumstances. Instead, it has been shown that implementing various types of feedback is a comprehensive way to improve students' accuracy and fluency while augmenting learner uptake and self-correction (Ellis, 2009; Lyster & Ranta, 1997). In particular, the evolving accuracy of Automated Writing Evaluation (AWE) has enabled automated feedback tools as assistants in the writing assessment process, addressing lower-level skills such as grammar while teachers devote more time to high-level writing skills such as content and organization (Link et al., 2020).

ChatGPT is an extensive language model developed by OpenAI (2024), which is designed for natural language processing tasks such as understanding text, generating human-like language, and even performing more complex tasks like analyzing and synthesizing information to provide coherent and contextually relevant responses. These functions make ChatGPT an effective automated tool that offers students feedback involving personalized comments on lexical, grammatical, and syntactical levels as well as suggestions for improvement based on specific writing needs. Consequently, the AWE functions of ChatGPT have been shown effective in supporting L2 learners with their writing by guiding them through corrections on grammatical errors and higher-order writing concerns like coherence and cohesion (Coyne et al., 2023; Yoon et al., 2023).

Several studies have explored the benefits of integrating ChatGPT into the L2 writing process, some others have determined the effectiveness of AI feedback in increasing students' text revision, motivation, and positive emotions (Meyer et al., 2024) as well as its overall impact as a feedback tool (Mahapatra, 2024). However, there is still very little empirical evidence regarding how the interaction with ChatGPT as an AWE tool may shape the language and considerations of appropriate peer feedback on ESL students, which is an important consideration now that AI tools are commonly used for automated feedback in ESL classrooms. This study is based on the hypothesis that students' language and criteria for providing peer feedback may be shaped by their interactions with ChatGPT based on ChatGPT's benefits for students' writing skills and positive emotions during automated feedback. Using a mixed methods approach, this study aims to advance the current knowledge of ChatGPT's role in the writing process and how it shapes students' language and perceptions of effective writing.

The methodological procedures and hypotheses presented in this study are based on a framework combining two theories: Vygotsky's (1978) Sociocultural Theory and Bandura's (1977) Social Learning Theory. Vygotsky's Sociocultural Theory proposes that learning occurs through communication and collaboration. This study looks at the Zone of Proximal Development (ZPD) as one of the mediators during peer and automated feedback interactions. According to the ZPD, social interactions provide scaffolding that helps students move from their current abilities to more advanced forms. Although this theory is not explicitly focused on language learning, Vygotsky's theory sets the foundation for many studies exploring peer-assisted learning in ESL and EFL contexts.

Bandura's Social Learning Theory (SLT) emphasizes individual learning through observation, imitation, and modeling. It explains that much of a person's learning occurs in social contexts where individuals can observe others' behaviors, understand the outcomes of those behaviors, and then replicate them. Through the SLT, it is possible to examine how students imitate peers or teachers in academic and social settings and, over time, internalize the observed behaviors. This indirect influence helps shape their language use, from vocabulary to speech patterns and pragmatics.

The indirect influence of ChatGPT on the language that students use to provide feedback was examined by analyzing linguistic features such as lexical diversity, sentence complexity, nominalization, modals, epistemic markers, and discourse markers. These linguistic features are relevant for comparing human and ChatGPT-generated language since these elements reveal differences in writing style, coherence, and complexity. Previous research has examined these features to assess writing quality and computationally model language use (Herbold et al., 2023; McNamara et al., 2010; Weiss et al., 2019). Lexical diversity, measured using the Measure of Textual Lexical Diversity (MTLD), indicates vocabulary richness; for instance, ChatGPT tends to use a broader range of synonyms compared to human writers. Sentence complexity, observed through syntactic dependency structures, shows how AI-generated texts often include embedded clauses, such as "Although the argument is valid, further evidence is necessary" (Herbold et al., 2023). In ChatGPT-generated language, nominalizations, like "improvement" instead of "to improve," are often included to contribute to a more formal style, while human-produced language uses more modal verbs ("should," "might") and epistemic markers ("I think," "it seems"), to express subjectivity. Discourse markers ("however," "therefore") are commonly used to provide a coherent sequence of arguments and ideas. The computational examination of these features provides a means of evaluating AI's impact on students' language use while providing feedback to their peers.

In order to test the hypothesis that students' language and criteria for peer feedback may be affected by the ZPD marked by their interactions with ChatGPT during the peer review stage of the writing process, this study aims to measure the effect of ChatGPT as an AWE tool on ESL students' language and criteria for providing peer feedback. This study was guided by the following research questions:

1. How does interacting with ChatGPT influence the linguistic features and evaluation criteria that students use when giving feedback to their peers?
2. To what extent does ChatGPT shape students' perspectives of helpful feedback?

Methods

Study Setting and Design

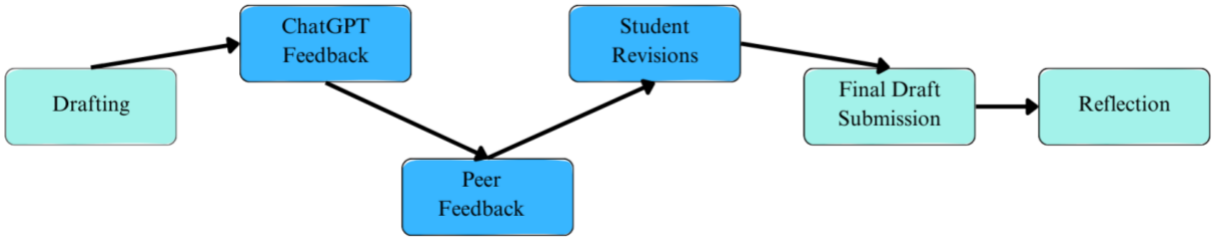
The investigation was conducted over five weeks with students in an asynchronous online English course at a university in the Midwest. The course was designed for international students to develop academic writing skills in English. Students were placed in the course based on their scores on an English Placement Test (EPT) assessing their proficiency in English for academic purposes. Therefore, it can be assumed that the participants in this research were proficient in using English as a second language but needed to develop their English for academic purposes. The peer feedback comments and ChatGPT interactions of eight students who volunteered to participate were used in the data collection and analysis in this research.

The study was conducted in the Fall of 2024 with approval obtained from the IRB. By then, OpenAI had enabled free public access to ChatGPT-4, the version that all students used for automated corrective feedback on their writing. Before starting the study procedures, students expressed high familiarity and previous experience with ChatGPT through a preliminary poll, so no instruction on ChatGPT use was needed before starting the procedures.

The procedures consisted of three writing tasks completed using a feedback loop. Each writing task asked students to write a different type of essay; for task 1, the participants wrote a fictional narrative; for task 2, they wrote a celebrity profile; and for task 3, they wrote a multimodal analysis speech. Each of these writing tasks was directly related to the course's major assignments so the tasks were relevant for the participants. Figure 5.1 illustrates the participants' writing process, including a feedback loop of automated and peer feedback.

Figure 5.1

Writing Process and Feedback Loop



While completing each task, the participants were asked to follow the following steps. First, they wrote the first drafts. Then, they asked ChatGPT to provide feedback on their drafts. In this step, the participants were shown examples of prompts asking for formal and global feedback. However, they used their own prompts according to their considerations of useful error correction. The participants were asked to export their conversations with ChatGPT and send them to the researcher for data analysis. Then, the participants provided feedback on two of their

peers' drafts on a Canvas discussion forum. The comments were required to be at least 180 words long. Participants were not given explicit instructions on what types of corrective feedback they should give. At the end of the feedback loop, the participants revised their drafts according to their peers' and ChatGPT's feedback and submitted the final drafts. After completing the three writing loops, they were asked to complete a survey about their perceptions of helpful feedback. In order to have an initial student language production benchmark without the influence of ChatGPT, the participants were asked to give peer feedback before interacting with ChatGPT during Task 1.

Collection of Feedback Corpora

The feedback comments from students and ChatGPT interactions were collected and compiled to create a student generated (SG) feedback corpus and a ChatGPT generated (CG) feedback corpus that each underwent linguistic analysis and comment-type coding. Each corpus was organized according to the writing task associated with the feedback. Each of the two resulting corpora were comprised of three subcorpora corresponding to Task 1, Task 2, and Task 3. The SG feedback corpus contained 697 sentences and 11793 tokens, while the CG feedback corpus contained 915 sentences and 19972 tokens. Then, all answers to the survey were collected for further analysis. Table 5.1 shows the size of the two corpora and their respective subcorpora.

Table 5.1

Student-Generated (SG) and ChatGPT-Generated (CG) Feedback Corpora

Corpora	Subcorpora	Sentences	Tokens
Student-Generated Feedback	Task 1	271	4126
	Task 2	201	3579
	Task 3	225	4088
Total		697	11793
ChatGPT-Generated Feedback	Task 1	328	6365
	Task 2	379	8580
	Task 3	208	5027
Total		915	19972

Data Analysis

Table 5.2 summarizes the analysis methods, the tools used to analyze the data, and the theories supporting the tools for each of the two research questions. To answer Research Question 1 about the influence of interacting with ChatGPT on the language complexity and evaluation criteria that students use when giving feedback to their peers, this study analyzed six main

linguistic features from the SG and CG feedback corpora. The analysis replicated the features and computations used to compare human-written and ChatGPT-generated essays in Herbold et al. (2023). In addition, the types of comments in SG feedback were coded and analyzed by the researcher.

Table 5.2

Methods Used to Address the Research Questions

Research Question	Analysis Method	Analysis Tool	Theory
RQ1	Linguistic Features	Computational Analysis	Herbold et al. (2023)
	Feedback Coding	Feedback Comment Type Human Coding	Rahimi (2013)
RQ2	Perception Degree	Likert Scale Elicitation	Likert (1932)
	Opinion	Open-ended Question Elicitation	

Using the linguistic complexity features proposed by Herbold et al. (2023), a longitudinal analysis of the language associated with the three tasks was conducted for both corpora. The occurrences were calculated using Python (Version 3.11), SpaCy (Honnibal & Montani, 2024) for natural language processing as well as Part-of-Speech tagging, and Pandas (Version 1.5.3) and Numpy (Version 1.24.1) for data analysis and computations. Sentence complexity was calculated by identifying and counting the syntactic dependency labels corresponding to adjectival clauses (e.g., "*The book that she recommended was great.*"), conjuncts (e.g., "*He studied hard, but he still failed.*"), adverbial clauses (e.g., "*She left because she was tired.*"), clausal complements (e.g., "*I share the belief that he is honest.*"), discourse elements (e.g., "Overall, the results are correct."), and parataxis (e.g., "*It was late; we decided to leave.*").

Lexical diversity was calculated using the MTLD, which identifies how frequently the vocabulary in a text is repeated regardless of overall text length. Higher scores indicate richer vocabulary. Scores above 90 suggest high diversity; below 60 suggests repetition or limited lexical range. The discourse markers score was calculated by identifying and counting the occurrences of the discourse markers defined in the Penn Discourse Treebank (Prasad et al., 2008) list of discourse markers, such as "*moreover,*" "*consequently,*" and "*on the other hand.*" Modals were identified and counted using the POS tagging module on SpaCy, including words like "*can,*" "*must,*" and "*ought to.*" Epistemic markers were identified using a list of epistemic expressions from Herbold et al. (2023), such as "*probably,*" "*I assume,*" and "*it is likely that.*" Finally, nominalizations were identified and counted using common nominalization suffixes such as -tion (e.g., "*organization*"), -ment (e.g., "*development*"), -ance (e.g., "*performance*"), and -ness (e.g., "*awareness*").

After computing the overall occurrences of the linguistic complexity features of both the CG and SG corpora, the means of each feature were calculated and normalized per 1000 words to ensure comparability of results. Lexical diversity values were the only linguistic features not normalized since the MTLD is length-insensitive and allows for accurate comparison regardless of text length differences.

To address the focus on student feedback criteria of Research Question 1, the current study replicated some of the methods for human coding used by Rahimi (2013), who examined whether training student reviewers could improve the quality of their feedback as well as their writing by analyzing students' comments on global aspects and formal aspects. The researcher did all the coding according to the following criteria. For coding the types of comments that students gave in their feedback to their peers, four categories were used: the formal and global categories, as used by Rahimi (2013), in addition to a category of appraisal and a category of stance, which were added for the purpose of this study. Examples of student language coded according to the four types of feedback and their corresponding criteria are shown in Table 5.3.

Table 5.3

Criteria for Coding Feedback Comment Types

Type of Feedback	Description	Student Feedback Example
Formal	Addresses surface-level aspects of writing, such as grammar, vocabulary, and mechanics. The focus is on the correctness of linguistic forms rather than the purpose or organization of the writing.	*Lastly, I think that here was a spelling mistake "He started packing his staff" Do you mean Stuff*
Global	Relates to higher-order concerns, including the overall content, organization, coherence, and development of ideas in the writing. It aims to improve the clarity and effectiveness of the writer's message.	*If you consider talking about a bit more on her father's actions, the readers can understand why he's behaving in this way*

Table 5.3 (continued)

Criteria for Coding Feedback Comment Types

Appraisal	Appraisal feedback focuses on positive reinforcement and recognition of strengths in a peer's writing. These comments aim to motivate and encourage the writer by highlighting aspects that were done well without suggesting changes or corrections.	*You gave your honest opinion, which is good and makes your speech stronger*
Stance	Stance feedback reflects the reviewer's personal opinions, attitudes, or perspectives on the content of the writing. It provides a subjective reaction rather than an evaluation or correction.	*It is funny that you are actually from Bahla, and every Omani know that Bahla is a scary place.*

After getting the counts for the linguistic features and feedback comment types for each of the tasks for each participant, mean counts for each of the three tasks across participants were calculated. A Friedman statistical test was applied to the counts of linguistic features as well as the comment code counts to determine whether there were significant differences in language features or types of comments across the writing tasks. Since the sample was small and the code counts were not normally distributed, the Friedman test was a good alternative to a repeated-measure ANOVA since it identifies differences across repeated measures for related samples for non-parametric data.

To answer research question 2, the perception of students about ChatGPT's and their peers' feedback was measured using a survey at the end of the data collection procedures. The survey contained two sections: ten questions using a Likert scale to quantify students' agreement with statements about the feedback process and four open questions where students could qualitatively express their perceptions.

Results

RQ1: How does Interacting with ChatGPT Influence the Linguistic Features and Evaluation Criteria that Students Use when Giving Feedback to their Peers?

Linguistic Analysis

Tables 4 and 5 display the linguistic feature analysis of student-generated language and ChatGPT-generated language across the three writing tasks. Each table shows the mean values normalized per 1000 words for each feature across tasks except for the lexical diversity values, which did not need to be normalized to ensure comparability.

Table 5.4

Normalized Mean Counts for Linguistic Features of Student-Generated Language

Feature	**Task 1**	**Task 2**	**Task 3**
Sentence Complexity	3.21	3.15	3.70
Lexical Diversity	79.86	95.11	79.52
Discourse Markers	0.92	0.70	1.03
Modals	0.13	0.04	0.13
Epistemic Markers	0.14	0.18	0.12
Nominalizations	0.84	1.43	1.73

Table 5.5

Normalized Mean Counts for Linguistic Features ChatGPT-Generated Language

Feature	**Task 1**	**Task 2**	**Task 3**
Sentence Complexity	3.44	4.18	2.45
Lexical Diversity	125.61	199.35	109.39
Discourse Markers	0.54	1.15	0.55
Modals	0.06	0.18	0.06
Epistemic Markers	0	0	0
Nominalizations	2.24	3.11	1.97

The analysis of linguistic features shows observable trends between student-generated and ChatGPT-generated feedback across the three tasks. In student feedback writing, sentence complexity and nominalizations have a progressive increase, suggesting gradual linguistic growth, while lexical diversity peaks in Task 2 but drops in Task 3. Discourse markers are used most in Task 3, and modal and epistemic markers are not too high but remain constant across the

tasks. In contrast, ChatGPT-generated feedback shows consistently high values for lexical diversity and nominalization, with Task 2 showing peaks in all the features, while epistemic markers are absent across the tasks.

Comparison of Student-Generated and ChatGPT-Generated Feedback Language

Table 5.6 displays the normalized means for each linguistic feature for SG and CG languages to provide a clear comparison between the two types of language production.

Table 5.6

Student-Generated and ChatGPT-Generated Linguistic Features Normalized Means

Linguistic Feature	Student-Generated	ChatGPT-Generated
Sentence Complexity	3.35	3.37
Lexical Diversity	84.83	118.8
Discourse Markers	0.88	0.71
Modals	0.10	0.10
Epistemic Markers	0.14	0
Nominalizations	1.33	2.41

A quick observation of the normalized means shows that ChatGPT produced more occurrences of linguistic features associated with formal and complex academic writing, like nominalizations, lexical diversity, and sentence complexity, while the students produced more occurrences of discourse markers, modals, and epistemic markers. ChatGPT produced higher lexical diversity (118.8 vs. 84.83) and a greater use of nominalizations (2.41 vs. 1.33), both of which contribute to a more abstract and information-dense writing style.

In contrast, student-generated texts showed more frequent rates of discourse markers (0.88 vs. 0.71) and epistemic markers (0.14 vs. 0.00), which indicates a more personally engaged tone. A higher occurrence of these features suggests that students are more likely to use language that guides the reader through their argument and reflects their stance on the content of the feedback. The differences in feature counts indicate that ChatGPT used more complex syntactic structures, likely mimicking structures related to academic or informational writing styles, while students showed a more moderate level of complexity that favored stance and personal expression. Figure 5.2 visually presents the differences between SG and CG linguistic features that were normalized per 1000 words. This figure excludes lexical diversity values since their relatively high numbers do not allow for an effective visual comparison on the scale used for the other metrics. Figure 5.3 presents the differences between SG and CG lexical diversity values.

Figure 5.2

Comparison of Student-Generated and ChatGPT-Generated Linguistic Features Normalized Means

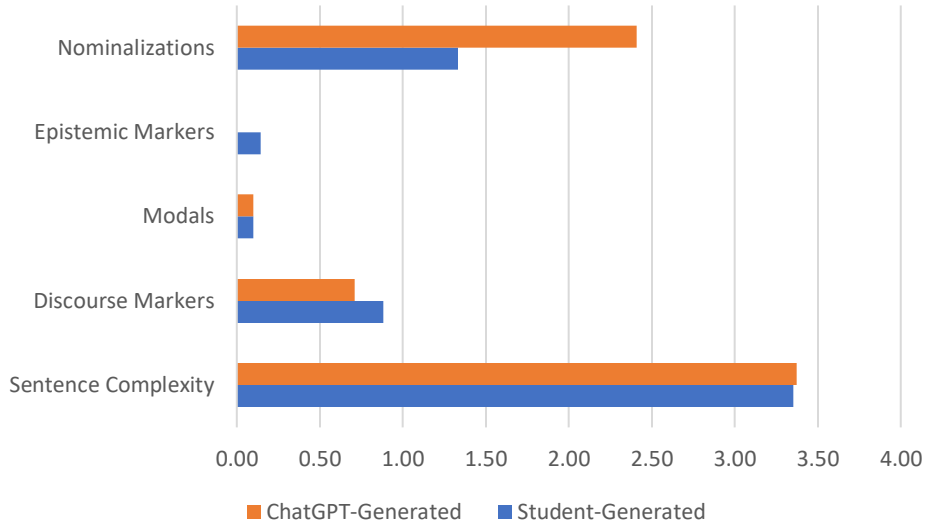
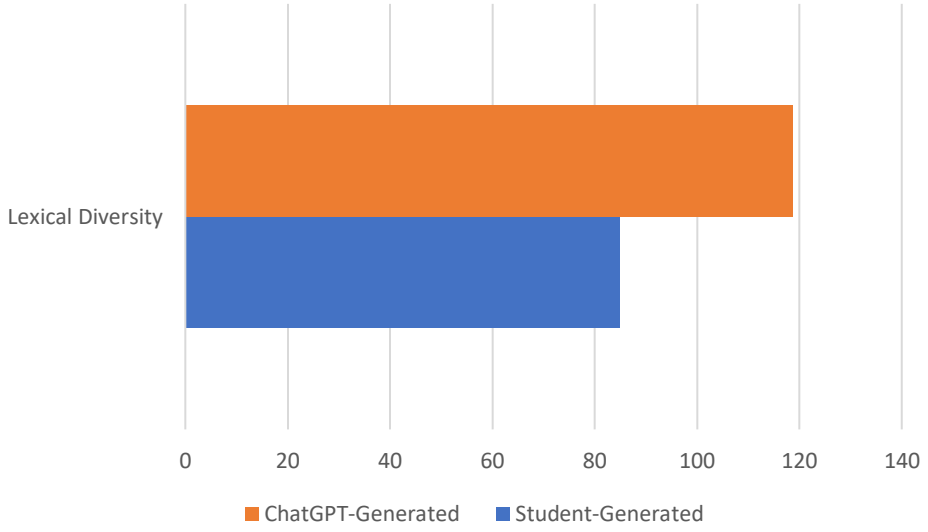


Figure 5.3

Comparison of Student-Generated and ChatGPT-Generated Lexical Diversity Measures



Both corpora show moderate use of discourse markers, but student-generated feedback has a slightly higher mean (0.88) than ChatGPT (0.71). Although this could suggest a more cohesive use of discourse markers by students, it is surprisingly similar in both corpora, which indicates that the most substantial differences between the two corpora were in the

morphosyntactic and complexity features, while the discursive use of cohesive devices is relatively similar.

The use of modals was the same in both student and ChatGPT feedback. However, ChatGPT's modal usage was concentrated on the second task, while students showed more balanced use across tasks, suggesting a constant descriptive writing style.

The most notable feature in SG language is the use of epistemic markers with a 0.14 occurrences per 1000 words, which, compared to 0 for ChatGPT, suggests that students expressed certainty or hedging more explicitly, whereas ChatGPT generated content with a more impersonal tone that is characteristic of its informational style used for stating facts.

ChatGPT used nominalizations extensively (2.41) compared to students (1.33), which reflects a more impersonal, informational style. Students show a gradual increase across tasks, but their overall usage remains much lower.

ChatGPT produced much higher lexical diversity, with a mean of 118.80, compared to 84.83 for students. These scores indicate that students repeat words more often, while ChatGPT incorporates a wide range of synonyms and varied word choices. However, it is important to note that although high lexical diversity can indicate fluency, excessive variation without contextual awareness may negatively affect coherence. In SG feedback, the participants exhibited less diversity, but their word choices appeared more deliberate and focused on their communicative intent. Examples of SG and CG language and their comparative descriptions can be observed in Table 5.7.

Table 5.7

Comparative Examples of Student-Generated and ChatGPT-Generated Language

Feature	ChatGPT	Student	Description
Sentence Complexity	*Instead of simply telling readers that Raneem is struggling financially and rethinking her decisions, show this through her actions and thoughts.*	*It also shows the strong bonding between Alex and Jamie and how they support one and other during hardtimes.*	The presence of several verbs in a sentence is a quick indicator of embedding. In one sentence, ChatGPT integrates several embedded clauses, a that-complement clause working as the direct object of a prepositional adverbial phrase, coordination, and a subordinated adverbial phrase beginning with "through". In the SG sentence, there are two coordinated clauses, and one of them is embedded in a wh-complement clause. However, sentence complexity in the CG sentence is notably higher than in the SG sentence.

Table 5.7 (continued)

Comparative Examples of Student-Generated and ChatGPT-Generated Language

Lexical Diversity	*Strengthen the ending by having Joanne reflect on what she learned, possibly imagining how her life might be different had she not experienced that night.*	*The plot was terrific, but you didn't explain the story entirely to the audience, especially the ending. I didn't quite understand what happened in the ending.*	These two passages provide similar feedback to the same text. ChatGPT uses varied vocabulary that avoids the repetition of lexical words. On the opposite, the student is emphatic on the fact that the ending was not clear. Although the SG sentence seems more repetitive, its word choice feels more intentional and emphatic than the one written by ChatGPT.
Discourse Markers	*The essay gives a solid view of Liam Gallagher and paints him as a resilient figure who embodies the spirit of Britpop. However, adding more examples and emotional depth would make it more compelling.*	*It is great that you have shared where and when he was born, and his achievements. However, there are some tinny grammar that could be better.*	The use of discourse markers was relatively similar between CG and SG language.
Modals	*Your introduction is effective, but to grab attention more firmly, you might try starting with a compelling question or statement.*	*If you were presenting that in front of people, you should use body language.*	Modals were similarly used by CG and SG to offer suggestions in an indirect manner.

Table 5.7 (continued)

Comparative Examples of Student-Generated and ChatGPT-Generated Language

Epistemic Markers	*N/A*	*I advise you to correct the spelling it seems that you have a lot of grammar mistakes as well.*	The use of expressions using subject pronouns combined with mental verbs (think, believe, know, guess, assume), verbs that express perceptions (feel, seem, look), and nouns and adjectives related to mental processes and emotions (belief, opinion; afraid, confident) are directly linked to expressing a stance. ChatGPT did not use any of these expressions in any of the tasks, while students used at least one of them per comment.
Nominalizations	*You've laid a solid foundation, but a few adjustments could improve its clarity, coherence, tone, and engagement.*	*I recommend proofreading your narrative to check conjugation problems. The Kevin's obsession with being the best and Julie's approach is relatable.*	The use of nominalizations was notably higher in CG language. While in the passage written by a student, there is one nominalization per sentence, the sentence written by ChatGPT contains five nominalizations in one sentence.

The Influence of ChatGPT Guidance on Student Feedback Language

In order to measure the significance of the linguistic feature differences across the three tasks, a Friedman statistical test was applied to determine which features showed significant changes that could not be attributed to chance. The p-value indicates the likelihood that these differences occurred purely by chance. P-values below the threshold (*0.1*) suggest that the observed differences are unlikely to be random and may be attributed to an underlying factor, which could be the influence of ChatGPT. Since this exploratory study uses a small sample size, the threshold p-value was set to *0.1* instead of the frequent *.05* to account for all possible significant trends that a smaller threshold could overlook.

Table 5.8 displays the results of a Friedman test applied to the linguistic features in the SG corpus across the three tasks. The features with a p-value below the *0.1* threshold show significant differences across the tasks that can be attributed to the influence of ChatGPT feedback interactions.

Table 5.8

Friedman Statistical Test of Student-Generated Feedback

Feature	p-value
Sentence Complexity	0.50
Lexical Diversity	0.09
Discourse Markers	0.58
Modals	0.09
Epistemic Markers	0.46
Nominalizations	0.04

With a p-value of *0.09*, lexical diversity and modals indicate slightly significant differences across tasks, possibly due to the influence of genre conventions and interactions with ChatGPT. Lexical diversity peaks in Task 2, likely due to the celebrity profile genre, which encourages topical vocabulary and an informational tone. Similarly, modals decreased in Task 2, suggesting that ChatGPT's influence and the descriptive nature of profile essays prompted a more direct and descriptive communicative style in students' feedback.

The p-value of *0.04* in nominalizations indicates that students' use of nominalizations varied noticeably across the tasks, suggesting an important significance in the increase of nominalization use. The number of nominalizations used by the participants increased steadily across the three tasks, suggesting that exposure to ChatGPT likely influenced students to use a more complex and academic vocabulary.

Indicators of Evaluation Criteria

Table 5.9 analyzes the types of feedback that students provided on their peers' writing across the three tasks, coded as to indicate how their evaluation criteria targeted formal aspects of language, referred to global aspects, displayed their appraisal, and indicated their stance.

Formal corrections decreased steadily across tasks, from an average of *2.12* in Task 1 to *0.62* in Task 3, which suggests that students became less focused on surface-level grammatical or lexical issues as the tasks progressed. Global corrections increased slightly from Task 1 (*3.12*) to Task 2 (*3.62*), then decreased to *2.87* in Task 3. This trend indicates that students initially focused more on broader writing aspects, but this focus declined slightly in Task 3. Appraisal comments showed a steady increase, from *3.87* in Task 1 to *4.62* in Task 2, reaching a peak of *8* in Task 3. This suggests that students progressively acknowledged positive elements in their peers' writing, emphasizing encouragement over error correction, especially in later tasks. Comments showing their stance decreased progressively across tasks, from *3* in Task 1 to *1.62* in

Task 3. This reduction suggests that students became less inclined to take positions or express personal beliefs about their peers' work over time, giving a more impersonal tone to their feedback.

Table 5.9

Types of Feedback per Comment Indicating Evaluation Criteria Across Tasks

Type of Feedback	Task 1	Task 2	Task 3
Formal	2.12	0.87	0.62
Global	3.12	3.62	2.87
Appraisal	3.87	4.62	8
Stance	3	2.12	1.62

Table 5.10 presents the results of a Friedman statistical test applied to the types of feedback coded across the three tasks in order to determine whether there were significant differences in the evaluation criteria students gave across the tasks.

Table 5.10

Friedman Statistical Test of Differences Among Evaluation Criteria Indicated in Students' Feedback

Type of Feedback	p-value
Formal	0.05
Global	0.78
Appraisal	0
Stance	0.08

A p-value of *0.05* in formal feedback indicates significant differences across the tasks, which suggests that the frequency of comments addressing grammatical or lexical errors decreased over time. This change demonstrates that students became less inclined to provide grammatical error correction to their peers after interacting with ChatGPT for feedback, while their global correction feedback showed no significant difference. A p-value of *0* in appraisal shows highly significant differences, which suggests a notable shift in students' use of appraisal comments, with a clear trend toward increased positive reinforcement or acknowledgment of strengths over time. This change can be likely due to ChatGPT's influence in positive reinforcement and explicit acknowledgment of good writing. A p-value of *0.08* in stance indicates slight significant differences that suggest that students' use of stance comments

decreased gradually, possibly linked to the influence of ChatGPT's impersonal and informational tone, leading students to take a more descriptive approach to writing feedback.

RQ2: To What Extent does ChatGPT Shape Students' Perspectives of Helpful Feedback?

Students' Perceived Agreement with ChatGPT's Influence

This question was addressed by describing the results of the two types of items on the questionnaire that students completed after participating in all three of the peer review tasks. The responses to the Likert-scale questions and the students' comments on the open-ended questions suggested that they saw value in ChatGPT's feedback.

Students' Likert Scale Responses

The highest-ranking scores (median=4 on the 5-point scale) indicating a positive view of ChatGPT's feedback appeared on questions 1, 6, and 10. Question 1 stated that ChatGPT's feedback helped students identify areas in their own writing that needed improvement; question 6 proposed that getting feedback from ChatGPT on their drafts influenced the way that they gave feedback to their partners; and question 10 asserted that ChatGPT was a good model to imitate for providing feedback or writing. The results from the linguistic feature analysis and the feedback criteria analysis support the perspectives that students expressed on the Likert scale, suggesting three possible interpretations. First, students found ChatGPT's feedback effective in highlighting weaknesses in their writing, which might indicate the tool's role in facilitating targeted revision. Second, students perceived ChatGPT's feedback as impactful in shaping their peer feedback practices, which suggests that interacting with ChatGPT extended their own capabilities for evaluating their peers' work. Third, students perceived ChatGPT as a valuable and reliable source for learning feedback techniques and writing strategies.

The questions receiving the least agreement were questions 5 and 7, both with a mean of 2.8 on the 5-point scale). Question 5 suggested that grammar correction is more important than feedback on content and structure; and question 7 proposed that when giving feedback to their peers, the participants used some of the evaluation criteria that ChatGPT used to evaluate their writing (like coherence, structure, word choice, etc.). The fact that these two statements were the lowest scored by the participants suggests two possible observations. First, students prioritized higher-order concerns like content and structure over surface-level grammar corrections, which aligns with the pedagogy in the course they were taking that teaches writing as a process focused on ideas and organization rather than solely on grammatical accuracy. The grammatical corrections provided by the technology did not overshadow the rest of the writing process for these students. Second, even though ChatGPT may have influenced students, they did not feel that they had explicitly imitated its evaluation criteria when giving feedback to their peers.

The Appendix A presents complete the results of the Likert scale section of the survey that students took after completing the three tasks. It contains the statement asked, the mean score (1 being the strongly disagree and 5 strongly agree), and the median score.

Students' Comments on ChatGPT's Feedback

The perceptions that students expressed in the open-ended questions asked in the survey about their interactions with ChatGPT for peer feedback were largely positive. Many participants emphasized ChatGPT's usefulness in identifying errors, improving clarity, and improving their writing style. One student mentioned ChatGPT's role in providing helpful feedback, which he defined as "feedback that actually helps me identify the places of error and help me improve my writing in a way that would make readers engage with my writing." Another student noted that "ChatGPT feedback improves my knowledge for error correction so that I could give appropriate advices in my friends." These responses suggest that students considered ChatGPT as a tool for receiving helpful feedback and as a means of developing their ability to give effective feedback to others.

Students also commented on how ChatGPT influenced their own writing. One participant explained that it "influenced my writing by improving word choices and giving some emotion or trying to engage readers in the best way." Others expressed similar viewpoints, describing ChatGPT's impact on the grammatical accuracy and organization of their feedback writing. For instance, one student shared: "I've learned from that way of writing with fewer grammatical mistakes." Another added, "ChatGPT often provides feedback that is concise and focused on making content clearer and more structured. This often helps me to give feedback with clarity to my peers." These statements suggest that ChatGPT served as a model that students followed in order to enhance their own writing in terms of clarity, organization and engagement.

However, not all participants shared the same level of enthusiasm. A few students expressed concerns about the limitations of ChatGPT's feedback. One participant stated, "To be honest, it didn't really help a lot. ChatGPTs feedback would make my writing less relatable. The writings will loose the human touch." This statement points out the concern that taking AI-generated feedback may diminish authenticity or personal style. Another student noted, "Even tho error correction is important to understand the writing, I felt ChatGPT was more focused on just pointing out errors and not giving an improvement feedback. And it as such didn't really influenced a lot because my feedbacks were more improvement based than just pointing out grammar." These comments suggest that some students viewed ChatGPT's feedback as surface-level corrections that lacked relatability, which highlights their preference for feedback and suggestions that help develop ideas and improve reader impact rather than localized corrections.

Discussion

The findings in this research study showed changes in students' language and evaluation criteria when they provide feedback to their peers using help from ChatGPT. The comparative analysis of student-generated and ChatGPT-generated language combined with the longitudinal analysis of students' linguistic features and type of feedback revealed four main findings. First, student- and ChatGPT-generated language differed in sentence complexity, lexical diversity, and epistemic markers, while it was relatively similar in its use of discourse markers. Second, interactions with ChatGPT for writing feedback influenced student peer review writing mainly in

three linguistic features: lexical diversity, modal verbs, and nominalizations. Third, ChatGPT influenced the types of comments that students gave to their peers when suggesting corrections. Their use of suggestions for formal grammatical correction and comments expressing attitude and opinion dropped, while the comments encouraging their peers' work increased notably. Finally, overall students' perceptions showed that they agree that ChatGPT is a good model for peer review and that it influenced their writing although this view was not unanimous.

Answering Research Question 1, findings indicated that interacting with ChatGPT influenced the linguistic complexity of students' feedback by increasing the number of nominalizations that they used while writing, which indicates a change toward a more lexically complex and academic style of writing. Students' lexical diversity peaked, and their use of modals decreased as they moved to the more informational genres, which could suggest a shift toward a more impersonal and descriptive writing style influenced by ChatGPT's informational discourse. Additionally, the change in the type of feedback students provided to their peers revealed a shift in evaluation criteria motivated by ChatGPT interactions. This change was highly noticeable in the increase in appraisal feedback comments and the decrease in formal correction and stance comments. This suggests a more descriptive and slightly impersonal writing style that favors encouragement and global aspects of error correction over grammatical and localized error correction. The observed changes align with Vygotsky's Zone of Proximal Development, which explains students' development of writing skills through scaffolded interactions, and Bandura's Social Learning Theory, which explains why students modeled some of ChatGPT's stylistic and linguistic patterns.

Answering Research Question 2, ChatGPT shaped students' perspectives of helpful feedback by influencing their feedback focus. This finding is supported by their perceived understanding of ChatGPT's role in their feedback process, as demonstrated by their answers to the survey. Survey responses indicate that students valued ChatGPT's feedback for its clarity and ability to identify areas for improvement, with the highest-rated statement being that ChatGPT highlighted weaknesses effectively and influenced their approach to giving feedback. However, questionnaire responses were mixed with not all students showing enthusiasm for ChatGPT's feedback.

Overall, this study aligns with prior research by demonstrating how AI used as AWE support enhances evaluative writing. Consistent with Weiss et al. (2019) and McNamara et al. (2010), the findings reinforce that AI-generated text exhibits greater syntactic complexity and lexical diversity (Herbold et al., 2023) compared to student-generated language. Additionally, similar to Rahimi (2013), it shows that structured support, whether from AI or human training, helps shift peer feedback from surface-level corrections to global evaluations. This chapter adds to prior knowledge by suggesting that the use of AI for automated-feedback can increase students' sentence and lexical complexity in addition to an increased focus on appraisal comments during peer evaluation.

Conclusion

This study provided evidence that a generative AI tool may help to mediate second language development by providing scaffolding within the learners' ZPD. The methodology included a pedagogical intervention that was used in a classroom setting like those in many institutions of higher education and intensive English programs, where students refine their second language writing skills in part through peer review. The qualitative analysis and descriptive data provide a granular view of how the GenAI tool affected learners' performance and perceptions as they participated in the intervention. The findings suggest that students' feedback language and evaluation criteria may have been influenced by their interactions with ChatGPT, but several limitations of the study should be considered when interpreting the results.

As an exploratory study, the results provide only an initial look at the influence of ChatGPT on students' peer feedback language and criteria. The interpretations are limited by the small sample size, lack of a control group, short duration of the study (five weeks). Additionally, all qualitative interpretations and coding were conducted by the researcher, leaving reliability unchecked. The researcher of this study also served as the course instructor, a dual role that contributed to the educational authenticity of the intervention while possibly introducing bias. Nevertheless, the findings suggest it would be worth continuing this line of research to continue exploring the possibilities for GenAI-assisted peer feedback on writing development.

References

Bandura, A. (1977). *Social learning theory*. Prentice Hall.

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). *Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction*. ArXiv. https://doi.org/10.48550/arXiv.2303.14342

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal*, *63*(2), 97-107. https://doi.org/10.1093/elt/ccn023

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617. https://doi.org/10.1038/s41598-023-45644-9

Honnibal, M., & Montani, I. (2024). spaCy (Version 3.7.5) [Software]. Explosion. https://spacy.io/

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19*(1), 37-66. https://doi.org/10.1017/S0272263197001034

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86. https://doi.org/10.1177/0741088309351547

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*(1), Article 9. https://doi.org/10.1186/s40561-024-00295-9

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

OpenAI. (2024). ChatGPT (Version 4) [Large language model]. https://openai.com/

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008). The Penn Discourse TreeBank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA). https://aclanthology.org/L08-1093/

Python Software Foundation. (2024). Python (Version 3.12.4) [Programming language]. https://www.python.org/

Rahimi, M. (2013). Is training student reviewers worth its while? A study of how training influences the quality of students' feedback and writing. *Language Teaching Research*, *17*(1), 67-89. https://doi.org/10.1177/1362168812459151

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Weiss, Z., Riemenschneider, A., Schröter, P., & Meurers, D. (2019). Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 30-45). https://doi.org/10.18653/v1/W19-4404

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023). *Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion.* ArXiv. https://doi.org/10.48550/arXiv.2310.06505

Appendix A

Table A1

Likert Scale Results

Number	Question	Mean	Median
1	ChatGPT's feedback helped me identify areas in my writing that needed improvement.	3.7	4
2	ChatGPT's feedback was aligned with what I consider good and helpful feedback.	3.5	4
3	The feedback I received from my peers helped me improve my writing.	3.6	3.5
4	The feedback I received from my peers was aligned with what I consider good and helpful feedback.	3.6	3.5
5	I think grammar correction is more important than feedback on content and structure.	2.8	2.5
6	Getting feedback from ChatGPT on my drafts influenced the way that I gave feedback to my partners.	3.8	4
7	When giving feedback to my peers, I used some of the evaluation criteria that ChatGPT used to evaluate my own writing (like coherence, structure, word choice, etc.)	2.8	3
8	When giving feedback to my peers, I used vocabulary words that I identified while receiving feedback from ChatGPT.	3.1	3
9	When giving feedback to my peers, I used discourse markers (overall, moreover, also, furthermore, etc.) that I identified when receiving feedback from ChatGPT.	3	3
10	I think ChatGPT is a good model to imitate for providing feedback or writing.	3.8	4

About the Author

Danilo Calle holds a TESL/Applied Linguistics Master's degree from Iowa State University. He has found his passion in teaching English as a means for community building and critical thinking development. Through his research, Danilo seeks to integrate his enthusiasm for corpus linguistics and computational linguistic methods with his passion for teaching to advance knowledge in language teaching and learning. He understands that the importance of applied linguistics research is reflected in the student-centered classroom, and he sees this as an opportunity to update his own teaching methodologies and constructs as well as to collaborate with his colleagues. Danilo has set a strong interest in researching the impact of artificial intelligence tools and digital spaces on the language development of ESL and EFL students.

Chapter 6

Empowering Youth Through Language, AI, and Sustainable Advocacy: A Telecollaborative Approach

Maria Mont[1], Melinda Dooly[1], and Onyemowo Angela Ejeme[2]
[1]*Universitat Autònoma de Barcelona*
[2]*Oakbridge Foundation*

This chapter explores the integration of Artificial Intelligence (AI) in language education through the lens of a telecollaborative project, "The Young Activists Project." Carried out over three months, the project involved 59 learners: 52 in Catalunya, Spain (ages 9-10 years old) and seven in Abuja, Nigeria (ages 10-14). This initiative combined AI with geographically distributed, project-based learning, with aims to enhance English language skills, digital literacy, and global awareness among young learners. A bespoke generative AI platform was utilized to facilitate language support for the multiple tasks in the project. Through multimodal interaction analysis applied to video recordings and screen recordings, this chapter examines if and how AI-driven tasks promoted language use and digital awareness during the project. Findings indicate that the young learners developed critical digital literacy skills by learning to refine their inputs for effective AI interactions, and they engaged with English more actively than in traditional classroom settings.

Cite as: Mont, M., Dooly, M., & Ejeme, O. A. (2025). Empowering youth through language, AI, and sustainable advocacy. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 103-121). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.06
©2025 Mont, Dooly, and Ejeme. Published under a CC BY license.

Introduction

In recent years, there has been considerable discussion about how Artificial Intelligence (AI) and digital tools are reshaping educational environments, including in the area of language teaching and learning. While AI tools are widely recognized for their benefits in language education, there is limited research on their impact on fostering language use and digital awareness. This gap highlights the need for age-appropriate AI tools and methodologies that address these specific challenges.

Second language (L2) learning and teaching is not exempt from the AI wave. Students have discovered that ChatGPT is quite adept at writing coherently in a variety of genres, providing substantive and well-organized answers to questions, and able to communicate in multiple languages. It is also capable of correcting grammar, improving writing style, and offering high-quality translation. “Teachers have discovered AI’s ability to create lesson plans; write texts or learning activities at any given level of proficiency; and provide assessments of student work” (Godwin-Jones, 2024, p. 6).

This same author argues that the role of the teacher is shifting and that AI represents "dynamically shifted ecosystems for language education" (Godwin-Jones, 2024, p. 8). While much of the discussion around AI in education focuses on its application within traditional classroom settings, researchers argue that true transformation requires rethinking pedagogical approaches (Celik et al. 2022). Godwin-Jones (2024) posits that language learning that takes place through "social interactions, such as peer collaboration and virtual exchange" (p. 15) will be enhanced by the use of AI resources.

Along these lines, this chapter examines a telecollaborative project entitled “The Young Activists Project”[1] which combines AI and geographically-distributed project based learning to foster language skills, digital literacy, and global awareness among young students. We refer to AI as the simulation of human intelligence in computers that are programmed to simulate human thinking, often using data-driven algorithms. Generative AI, which encompasses artificial intelligence systems capable of generating original content such as text or speech, is utilized in this project to facilitate real-time, contextually relevant language interactions. Furthermore, we have employed this technology to develop customized instructional materials tailored to their students' specific learning needs. In our case specifically, a bespoke AI-driven language model was tailored specifically for children under 13 years old, with privacy and regulatory compliance (COPPA, GDPR) as a priority. Through the dynamic AI-platform students used the target language (English) for real-time responses related to the telecollaborative project.

Studies demonstrate that AI offers significant benefits for language education, including intelligent tutoring systems, automated feedback, and tools that foster multilingualism and autonomy (Chen et al., 2020; Liang et al., 2021). Research also shows that these tools can enhance fluency, personalize learning, and support innovative teaching practices (Moulieswaran & Prasantha Kumar, 2023). Nonetheless, research also highlights gaps, particularly in areas like critical thinking, collaborative learning, and skills for young learners (Liang et al., 2021). These

[1] (https://sites.google.com/xtec.cat/the-young-activists-adventure/home)

gaps underscore the importance of exploring how AI tools can be effectively integrated into pedagogical approaches that prioritize engagement and age-appropriate content for young learners.

Although there is less research on AI and young learners, it is recognized that young children are increasingly interacting with AI applications in their daily lives. Prior research has demonstrated that there is a growing use of AI tools in early childhood education (ECE) settings, and that its use holds potential benefits (e.g., Lin et al., 2020; Tseng et al., 2021; Williams et al., 2019). Recent studies have introduced the concept of "AI literacy" to emphasize the necessity of incorporating AI into the suite of 21st-century digital literacy skills applicable to all individuals, including young children, although some scholars argue that insufficient attention has been directed toward fostering AI literacy and understanding its associated impacts on younger learners (Su et al., 2023). As Woo and Choi (2021) explain, many studies emphasize AI's role in personalizing language learning; fewer studies focus on its potential for language learning or critical digital literacy among children.

Still, some studies suggest that integrating AI tools can help children understand AI concepts, practices, and perspectives (Durrani et al., 2024; Su et al., 2023). However, challenges such as inadequate curriculum design, limited teacher expertise, and a lack of teaching guidelines persist (Su et al., 2023). Also, as Gutiérrez (2023) points out, while these tools may enhance engagement and literacy, concerns about their appropriate use, accuracy, and cultural biases remain. Some scholars may question whether early primary students are too young to engage with and acquire AI-related knowledge (Su et al., 2023). Research on digital literacy suggests that young learners often lack the skills needed to critically evaluate AI-generated content, a challenge that requires targeted educational interventions (Su et al., 2023). This chapter aims to address these gaps by investigating how AI can enhance foreign language use and digital awareness in a telecollaborative setting. This chapter analyses data taken from one of the schools involved in the telecollaborative project to answer the following research questions:

1. Did the AI-driven telecollaborative project promote the use of the target language (English) with the young learners? If so, how?
2. Did the use of AI-enhanced tasks promote digital awareness in the young learners? If so, how?

Methods

Project Context and Participants

The project was conducted over a three-month period and structured into distinct phases. A total of 63 participants took part in the study, comprising 52 learners aged 9–10 years in Catalonia, Spain, and seven learners aged 10–14 years in Abuja, Nigeria. Additionally, two teachers—one from each partner school—collaborated with two researchers in the design and implementation of the telecollaborative project.

The project focused on Sustainable Development Goal 6 (SDG6), which is part of the United Nations' 2030 Agenda for Sustainable Development. This goal aims to ensure the

availability and sustainable management of water and sanitation for all. The primary objectives of the project were to:

- Develop students' language proficiency through meaningful interaction with AI and peers in English.
- Enhance digital literacy by teaching students how to effectively interact with AI platforms.
- Cultivate global awareness and a sense of responsibility by engaging students in discussions about water conservation and sustainability.

Through a structured yet flexible curriculum, students were encouraged to interact with AI tools, engage in collaborative tasks, and produce creative outputs. By embedding the project within the framework of SDG6, educators were able to contextualize language learning in a real-world problem, which, as research suggests, can increase student motivation and engagement.

The project was divided into phases that built progressively on students' language, digital, and cultural competencies. Drawing on prior research in telecollaborative language learning (Dooly, 2008), this timeline provided the necessary scaffolding for students to develop complex skills and tasks.

1. **Preparation and Orientation**: Students were introduced to basic AI functions, digital literacy principles, and the water conservation topic. Initial sessions emphasized building foundational knowledge and familiarizing students with the AI tools and writing prompts.
2. **AI-Enhanced Language Learning**: Students engaged in structured language tasks using a customized ChatGPT platform tailored to young learners' needs, fostering language skills through interactive and AI-driven prompts. The ChatGPT platform was customized to meet the specific needs of young learners by restricting its language capabilities to English only, ensuring that Spanish-speaking students had to communicate exclusively in the target language. Additionally, it was pre-loaded with topic-specific prompts related to the project's focus on Sustainable Development Goal six (SDG6) concerning water and sanitation. These customizations provided structured yet interactive opportunities for students to engage meaningfully with both the topic and the target language. Consent for AI usage and data analysis (as a part of two Ph.D. research projects) was obtained from all participating families.
3. **Global Awareness Activities**: Focusing on local and global water issues, students collaborated across borders, discussing and researching topics, and developing a play that highlighted water conservation.
4. **Reflection and Final Presentation beyond their classroom walls**: Students prepared final presentations in various formats (e.g., digital storytelling, theatre plays, songs and choreographies, comic strips, creating handmade posters, infographics…) to technologically share their learning with a broader audience, thus reinforcing their understanding and commitment to sustainable practices.

The project utilized various digital tools, which were carefully selected to ensure accessibility, functionality, and alignment with project goals. The principal one was the aforementioned customized ChatGPT platform which was tailored to ensure the interaction was age appropriate and responded to English only. This was complemented by several other digital tools (e.g., Canva, Google Meet, Plickers), some of which also incorporated AI (e.g., Invideo.io, Magic School AI).

Data Compilation

Data were gathered through video-recordings of the sessions (22 sessions of 1 hour each), along with screen captures of the students' use of the AI platform (10 hours of recordings). These video-recordings reflect in-class interactions (e.g., pair or group work in Spain). There are also video-recordings of the video conferences held between the two classes (3 and a half hours of recordings). In this case, the interactions are between the countries (class to class; the ratio in this case was 52 to 7). Students from each class came up one at a time to ask a question (Figure 6.1). In the case of Spain, the teacher selected the students to speak, whereas in Nigeria, all the students had the opportunity to interact with their international peers.

Figure 6.1

Video Conference between Classes



The recordings capture both the oral use of the target language (speaking to each other and the teacher during group work) as well as written target use while they are typing into the platform; this was captured through cameras focusing on the screens. The second question was "Did the use of AI-enhanced tasks promote digital awareness in the young learners?" The video-recordings were analyzed through the lens of learners' engagement with the AI-platform, and

analysis of the discussion and oral interactions focused on features of the platform, and the learners' ability to perform (and adapt performance) to complete assigned tasks through the use of prompts. The video-recordings of conversations and students' interactions with the screens provide direct evidence of language use, engagement patterns, and digital behaviors, allowing researchers to analyze how the AI-driven telecollaborative project influenced the young learners' target language use and digital awareness.

The analysis of the data was carried out within the framework of the Autonomous University of Barcelona's doctoral program, which includes an annual review process addressing data management and ethical considerations. The research adheres strictly to the university's Code of Good Practice, which aligns with the ethical principles outlined in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), ensuring respect for persons, beneficence, and justice.

Before collecting the data, parental consent was obtained. The parents were invited to an informative meeting where the research project was presented and all questions related to the collection and use of data were answered. To ensure their anonymity, pseudonyms have been used in this study and the video stills (images) have been blurred (according to their consent forms).

Analytical Approach

In order to answer the research question "Did AI-driven telecollaborative project promote the use of the target language (English) with the young learners?", observational, multimodal interaction analysis was applied to video recordings and screen recordings. The analysis adhered to the central principles of Conversation Analysis (Seedhouse, 2009), in which participants' actions are tracked sequentially (in multiple modalities) through an emic (participant-oriented) lens.
The recordings were first transcribed before being segmented into smaller datasets for analysis of target (and L1) language use. This allows for observation of recurring patterns of social interaction that clearly illustrate the phenomenon under study. Fragments that fully encapsulate the interactional feature being analyzed are organized to be analyzed in more detail. For instance, in this study, the segments were analyzed for different aspects of language use (e.g., negotiations for meaning, or language used for specific functions such as giving commands, responding to cues from AI). Their language use is then compared with communicative intentions. Similarly, learners' performance, indicating their ability to produce correct prompts and understand AI responses, was analyzed for digital literacy and awareness development.

Results

We divide this section into two main parts. First, we consider evidence to answer RQ1 before looking at the data in regard to RQ2. Following the protocols of Conversation Analysis (Seedhouse, 2009), the fragments included here were chosen based on their relevance to the research question, as they demonstrate how social actions are accomplished through talk.

Regarding RQ1, "Did the AI-driven telecollaborative project promote the use of target language (English) with the young learners?" there is evidence that the young learners' engagement with the AI platform stimulated direct responses in the target language. There is also evidence of comprehension in students' responses and translations of texts as they engaged with the AI responses.

In Fragments 1 and 2, the students are working in pairs or trios with colleagues from their class with the bespoke ChatGPT platform. Their instructions are to use the platform "to get smart, yes-no questions that will help them guess the country of their partners". The transcription begins at 50 seconds into the recording. The transcription key can be found in Appendix A.

Fragment 1. *Comencem*?

S1:	**comencem?**
Trans:	*do we start?*
Tchr:	of course\ (.) yes\ (.) >ready steady GO/<
	((typing noises; students discuss in Catalan how to do the activity))
ST INPUT ONSCREEN:	hello
S1:	**vale\ te envio**
Trans:	*ok\ i send it to you*
S2:	**vale\ pero ponlo aquí**
Trans:	*ok\ but put it here*
S2:	**solo que-**
Trans:	*It's just that-*
ChatGPT INPUT ONSCREEN:	Hello! How are you today? If you have any questions or something fun you want to talk about, just let me know! [smiley face emoji]
S2:	((reading in low voice)) hello (.) how are you
S1:	**vale\ quita la mesa y ponte allí**
Trans:	*ok\ move the table and sit there*
S2:	[hello\ how are you today/ if you have any questions]
Tchr:	[do you remember the questions/]
S1:	**espera (.) lo voy a poner-**

Trans:	*wait (.) i'm going to put-*
S2:	i'm fine\
S1:	no no (.) **lo voy a poner** (.) are you good? (.) no\ i'm good\ (.) and you?
Trans:	*i'm going to put*
S2:	**vale**
Trans:	*ok*
S1:	((typing and speaking)) i am good (.) and you?
S2:	**pero hay que poner eso?**
Trans:	*but are we supposed to put that?*

We can see in this fragment how the learners discuss and negotiate how to interact with ChatGPT. While this brief discussion is carried out principally in Catalan and Spanish, the students provide input to the AI platform directly in English, demonstrating how the functionality and purpose can have an impact on the task outcomes (Casal & Bikowski, 2020). We can also observe how the young learners notice and respond to the social cues of the AI-prompts, corroborating other studies that indicate that the 'human-like' quality of the technology such as politeness may have an impact on participant engagement (Dynel, 2023).

In the next fragment (2), which takes place at minute 12:52, the students are still trying to come up with good yes-no questions to ask their partners in the upcoming videoconference.

Fragment 2. ***Are you from Africa***

ST INPUT ONSCREEN:	Are you from Africa
ChatGPT INPUT ONSCREEN:	I'm not from Africa or any other place because I'm a virtual assistant created by OpenAI. But I can share lots of information about Africa or help with questions about it. What would you like to know? (smiley face emoji)
ST INPUT ONSCREEN:	hello
ChatGPT INPUT ONSCREEN:	Hello again! How can I help you today? (smiley face emoji)
S2:	hello again\ **dice otra vez como te puedo ayudar/**
S1:	**jo tío** ((typing: we))
Trans:	*jeez man*
S1:	((speaking and typing simultaneously)) we are pl::a::ying (.) the mi-

S2:	**<u>alà (.) es muy servicial\</u>** (.) ((reading)) how can i help you today/
Trans:	*wow (.) it's very obsequious*
S1:	**<u>a mi</u>** me XXXX
Trans:	*to me too XXXX*
S1:	((speaking and typing simultaneously)) we are playing (.) the mistery skype\ ((types in mistert then corrects the spelling))

At this point, we can observe that they are able to adjust their prompts in the target language (Xiao & Zhi, 2023) as they seem to have become more adept at giving the prompts (explaining the context of what they need), although they are still unable to get the responses they need. As with the first fragment, we can see they not only read the responses (which can be used by teachers to practice pronunciation) but they are also dictating what they are typing into the platform, and thus further practicing their pronunciation. In both fragments, we see evidence that the learners are comprehending what they are reading through their discussion and interpretation of the written input (Pegrum, 2014).

We now turn to RQ2: Did the use of AI-enhanced tasks promote digital awareness in the young learners? If so, how? In the same context as the previous examples, a different group of students were asked to use AI for the very first time. They were instructed to use the AI platform to get help to play the Mystery classroom game (goal: come up with 20 yes/no questions). They were reminded that the questions must be smart and strategic to be able to get as much information as possible to successfully discover the secret location of their partner school.

Figure 6.2 (below) is an example of interaction with the personalized ChatGPT. The screenshot is taken at 2 minutes, 15 seconds into the students' interaction with the platform. We can see their prompts do not generate the answers they needed, because their prompts were not specific enough.

Approximately five minutes on, at minute 7 20", the same group of students adjusted their interactions to supply more specific and concise prompts (see Figure 6.3 below). However, they were still lacking some relevant information, such as questions that must be yes /no. So, once more, the AI answers were not yet satisfactory (the answers were long and dialogic in nature).

Figure 6.2

Example of Interaction with the Personalized ChatGPT

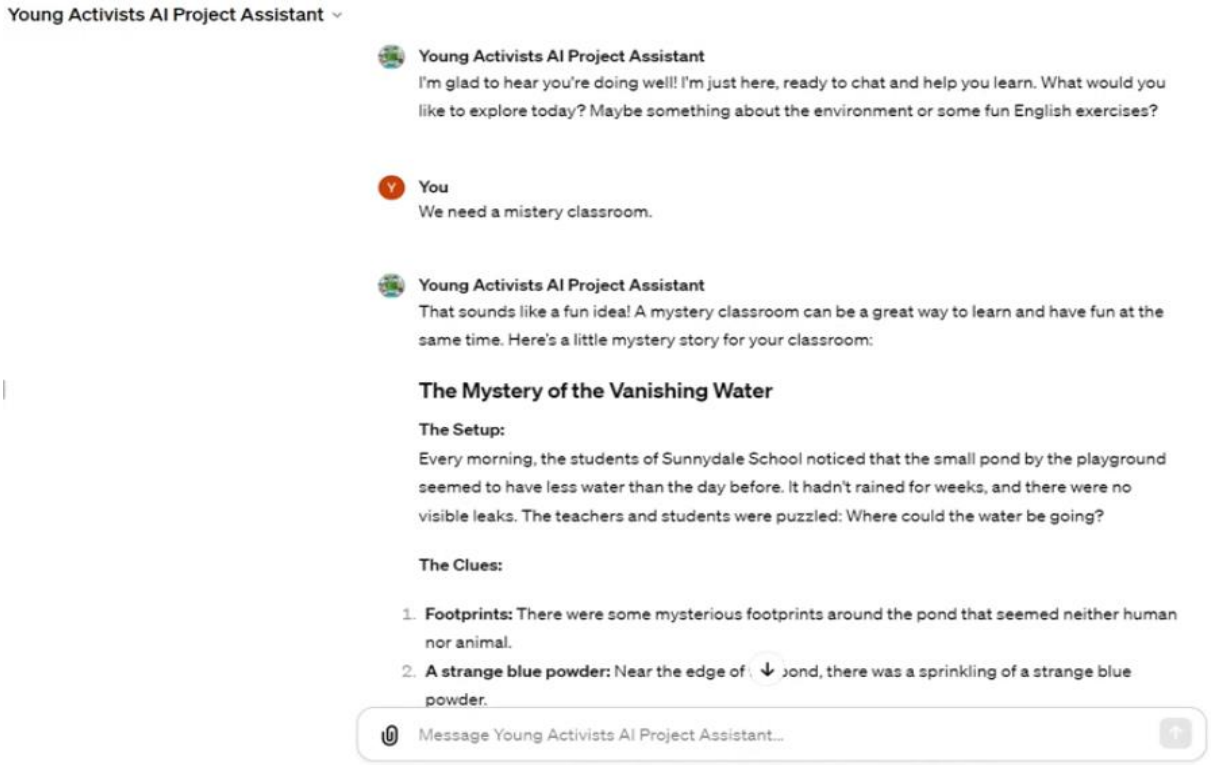
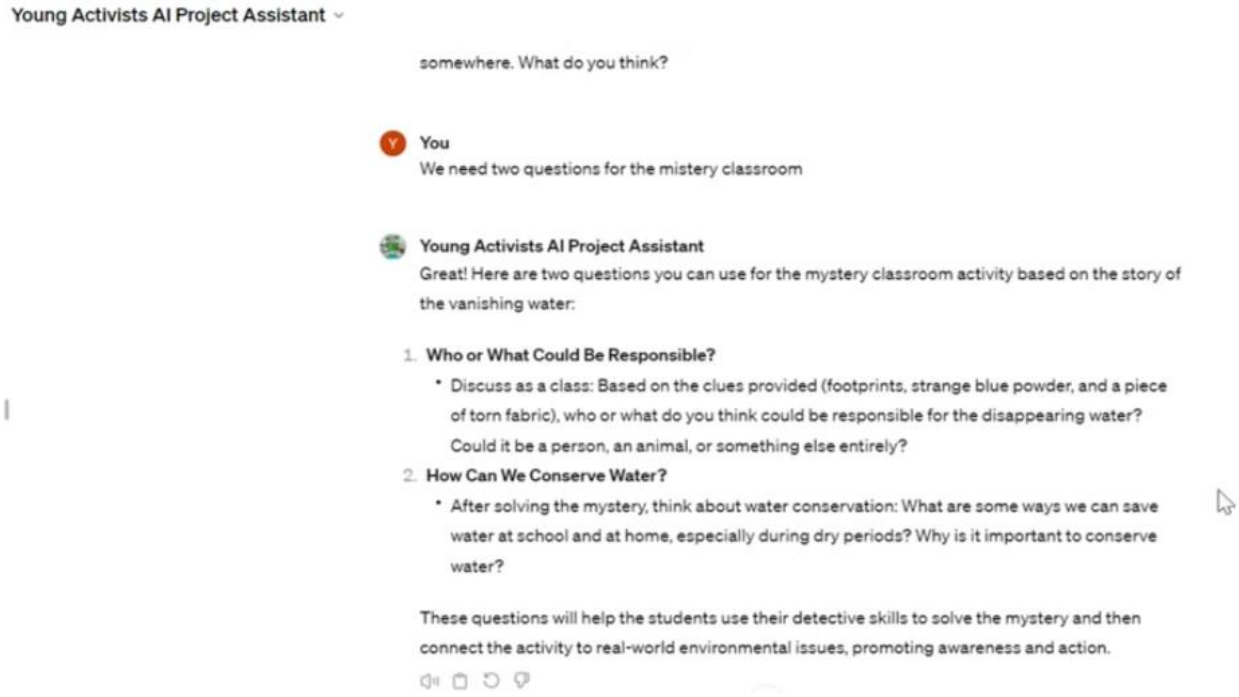


Figure 6.3

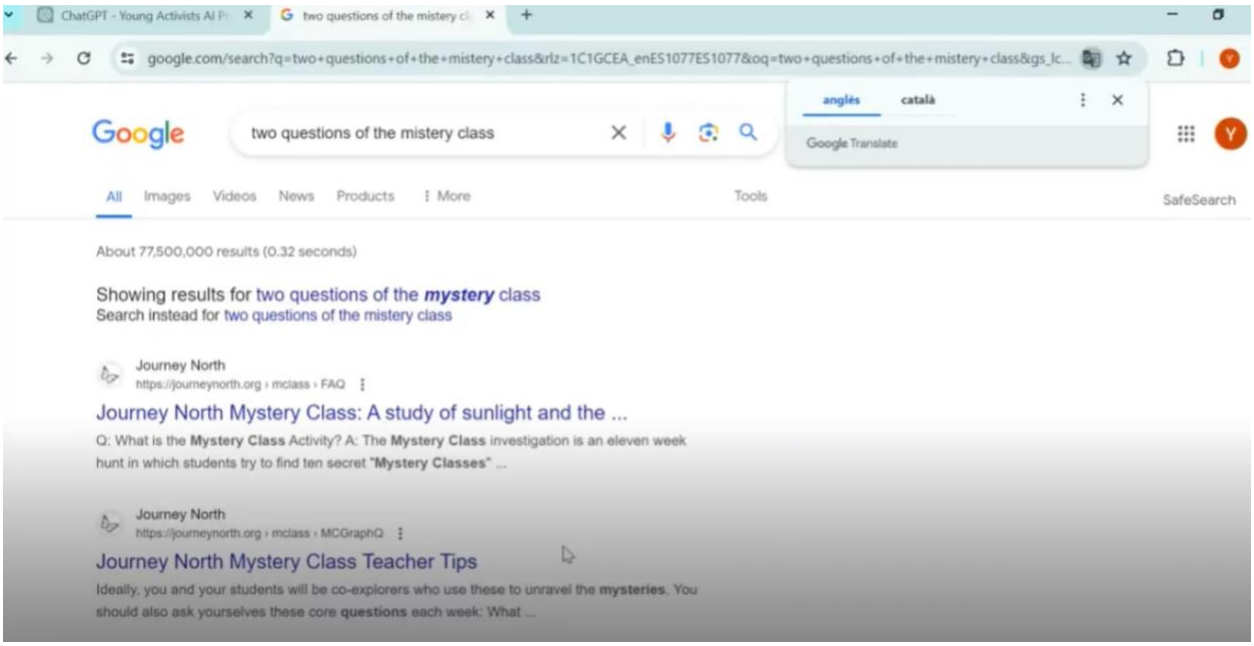
Example Showing Students' Adjustments toward More Specific and Concise Prompts



A bit further on, at minute 9 52” the students became frustrated with AI and the responses they were getting and turned to Google as a presumably better resource (Figure 6.4). However, they realized immediately that the answers were not even close to what they wanted and needed (Google gave them expansive examples of Mystery Class projects).

Figure 6.4

Screen Showing Students’ Interaction with Google



This prompted further discussion and negotiation—primarily conducted in Spanish and Catalan—indicating that the use of AI promotes digital awareness regarding how they must provide input. Finding that the Google search engine did not provide them with the answers they needed, the students returned to the AI platform and gave examples of ‘intelligent questions’ (Fragment 3, below; minute 22’ 42”). At this point, the students are only partially successful in providing good prompts, although they are beginning to provide more specific details in their descriptions.

Fragment 3. my Cuestion thats intelligent

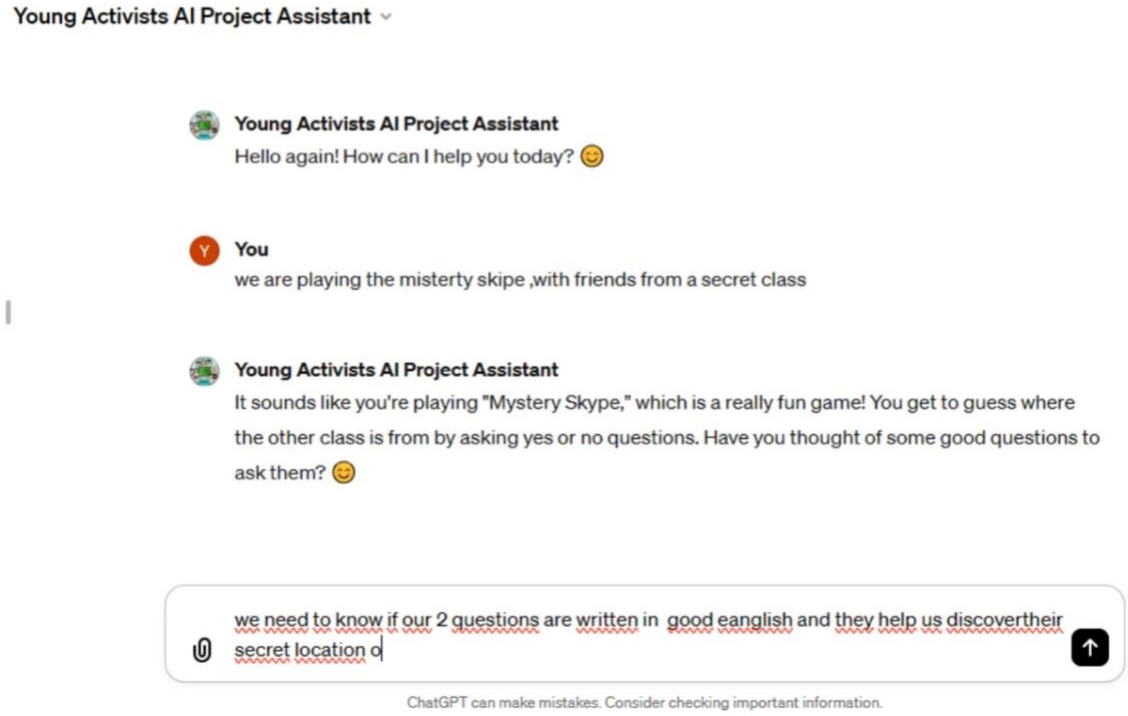
ST INPUT ONSCREEN:	“my Cuestion thats intelligent: are you from”
S1	**es** QUESTION (.)\ question **con Q**
Trans	*It is “Question”, Question”, with Q*
S1	**no: con Q**
Trans	*No, with Q*

ST INPUT ON SCREEN	“my Questions thats intelligent: are you from”
S2	my questions that’s intelligent (.) are you from America?
ST INPUT ON SCREEN	“my Questions thats intelligent: are you from America?”

Through the examples above, participants showcased enhanced interaction skills with AI, beginning to learn to craft effective prompts during their initial sessions. The following analysis (Figure 6.5) illustrates how their interactions with ChatGPT became increasingly precise and successful, enabling them to achieve their desired outcomes while using English as the primary medium of communication with the tool.

Figure 6.5

Example Showing Enhanced Precision and Success in Student–ChatGPT Exchanges

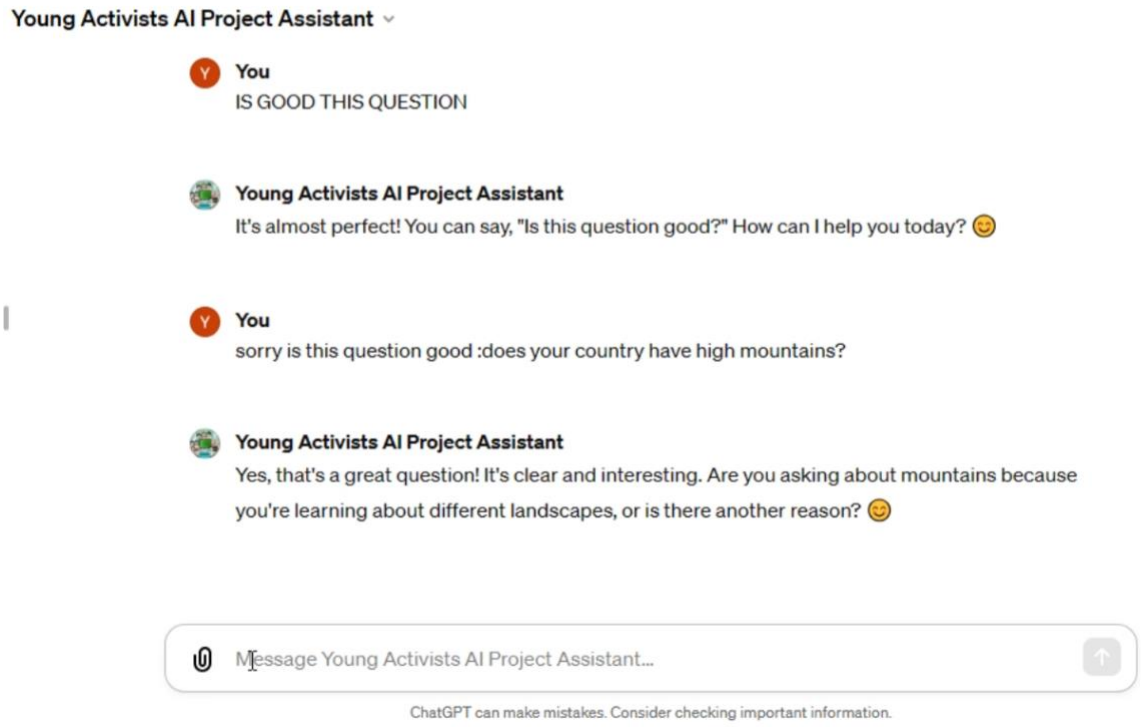


Comparing the input shown in Figure 6.2 to the input on the screen in Figure 6.5, we can see that the students transition from stating 'we need a mystery classroom' and searching on Google, to providing a relatively long prompt that deliberately requests feedback on the questions they have written (Figure 6.5): “we need to know if our 2 questions are written in good english [sic] and they help us discover their [sic] secret location.

And as is evidenced in Figure 6.6 (below) the AI platform corrects their English use “you can say, “Is this question good” in response to the prompt “IS GOOD THIS QUESTION”. To which, there is not only uptake on the correction, but the student responds to the error politely: sorry is this question good: does your country have high mountains?”. This brings back into play the pragmatic gains discussed in regard to RQ1.

Figure 6.6

Example of Students' Uptake of ChatGPT's Correction and Demonstrated Pragmatics Gains



Discussion

In summary, as students worked with AI, they improved skills in evaluating digital responses, an important component of digital literacy and today's AI world. Early interactions often led to frustrations, with students comparing AI responses to those of more familiar search engines like Google. This comparison process helped students identify AI's unique strengths and limitations, contributing to their critical digital literacy development and everything autonomously, without the help or guidance of their teachers. Students learned that AI, while capable of generating responses, required clear, specific input to function effectively, underscoring the collaborative nature of human-AI interaction (Griffiths, 2023).

The example above demonstrates that students noted that the AI's responses improved markedly with direct, structured prompts. Through practice, students learned to "feed" AI more accurate information, gaining confidence in their ability to guide digital tools toward meaningful interaction. This aligns with interactionist theory in second language acquisition, which emphasizes interaction as fundamental to language learning (Chapelle, 2001), extending to interactions with AI as collaborative partners. This approach reflects a synergistic framework where human and AI intelligence combine to form collaborative intelligence (Griffiths, 2023).

The Young Activists Project has demonstrated the potential of AI in fostering language learning. The evidence of comprehension in students' responses and their ability to translate texts demonstrates that the learners not only used the target language to interact with the platform, but

they also understood and processed the information it provided. Moreover, the AI-driven telecollaborative project appears to have encouraged the young learners to engage in the target language (English), potentially more than they would in a regular classroom setting. We also see that the students notice and respond to the social cues of the AI prompts, providing teaching opportunities related to language pragmatics. Despite these positive gains, it must be noted that while there were opportunities for oral language practice between the partner learners, the lack of feedback makes it challenging to assess effectiveness, highlighting a gap in research on pronunciation practice through AI. More research is needed to explore these areas further.

The Young Activists Project has also demonstrated the potential of AI in fostering digital literacy. One of the most significant areas of digital literacy growth was evident in how students refined their inputs for the AI platform. Initially, their prompts were vague or overly complex, which led to unsatisfactory responses. Through trial and error, students quickly learned to craft more specific and contextually relevant prompts, thereby achieving more effective outcomes. This iterative process exemplifies the principles of the "learning by doing" paradigm (Kolb, 1984), wherein hands-on engagement and experimentation are central to skill acquisition. By navigating the challenges and complexities of AI interactions, students engaged in experiential learning that mirrored authentic, real-world scenarios. Such engagement underscores the broader educational value of integrating AI tools into learning environments, as it equips students with critical 21st-century competencies, including problem-solving, adaptability, and digital fluency.

Conclusion

The Young Activists Project has illuminated the transformative potential of integrating AI-driven telecollaboration in language learning and digital literacy education. By embedding language practice within the context of Sustainable Development Goal 6 (SDG6), the project effectively bridged linguistic and cultural gaps between students from two different backgrounds. Findings reveal that the AI-enhanced tasks not only promoted the use of the target language (English) but also fostered students' digital awareness and critical thinking. Through iterative interaction with AI platforms, learners developed refined prompting skills, embraced the "learning by doing" paradigm, and demonstrated resilience in adapting to novel technologies.

The Young Activists Project has demonstrated the potential of AI in fostering language learning, digital literacy, and global awareness. By enabling students to engage in real-world tasks with peers from across the world, the project nurtured a generation of young activists—students who, we hope, will carry forward their sense of environmental responsibility, critical awareness, and collaboration well into their future.

Future research should explore the longitudinal impacts of AI-assisted language learning on fluency and retention, as well as the evolving role of teachers in these dynamic learning ecosystems. Additionally, there is a need to investigate the integration of feedback mechanisms for improving oral proficiency and critical digital literacy. The Young Activists Project serves as a promising blueprint for leveraging AI to empower young learners, fostering not only linguistic

and digital competencies but also a deep-seated commitment to addressing global sustainability challenges.

Acknowledgments

This work has been carried out within the framework of the doctoral program in Education at the Autonomous University of Barcelona.

We would like to express our sincere gratitude to GREIP Research Group and Ese Emmanuel Uwosomah for his invaluable contributions to this project and study.

References

Casal, J. E., & Bikowski, D. (2020). A framework for learning with digital resources. Applications for project-based language learning. In G. H. Beckett & T. Slater (Eds.), *Global perspectives on project-based language learning, teaching and assessment: Key approaches, technology tools and frameworks* (pp. 167–184). Routledge. https://doi.org/10.4324/9780429435096

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *Tech Trends, 66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research.* Cambridge University Press. https://doi.org/10.1017/CBO9781139524681

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8,* 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Dooly, M. (Ed.). (2008). *Telecollaborative language learning: Moderating intercultural collaboration and language learning. A guidebook to moderating intercultural collaboration online.* Peter Lang. https://doi.org/10.3726/978-3-0351-0565-0

Durrani, R., Iqbal, A., & Akram, H. (2024). Artificial intelligence (AI) in early childhood education, exploring challenges, opportunities, and future directions: A scoping review. *Qlantic Journal of Social Sciences, 5*(2), 411–423. https://doi.org/10.55737/qjss.135537445

Dynel, M. (2023). The role of politeness in human-computer interactions in language learning: A qualitative study. *Journal of Pragmatics, 208,* 1–13. https://doi.org/10.1016/j.langcom.2023.09.002

Godwin-Jones, R. (2024). Distributed agency in language learning and teaching through generative AI. *Language Learning & Technology, 28*(2), 4–31. https://doi.org/10.64152/10125/73570

Griffiths, C. (2023, June 19). *ChatGPT elevates the need for critical thinking skills*. ERE. https://perma.cc/JC54-SUP7

Gutiérrez, L. (2023). Artificial intelligence in language education: Navigating the potential and challenges of chatbots and NLP. *Research Studies in English Language Teaching and Learning, 1,* 180–191. https://doi.org/10.62583/rseltl.v1i3.44

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development.* Prentice-Hall.

Liang, J. C., Hwang, G. J., Chen, M. R. A., & Darmawansah, D. (2021). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments*, *31*(7), 4270–4296. https://doi.org/10.1080/10494820.2021.1958348

Lin, P., Van Brummelen, J., Lukin, G., Williams, R., & Breazeal, C. (2020). Zhorai: Designing a conversational agent for children to explore machine learning concepts. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), 13381-13388. https://doi.org/10.1609/aaai.v34i09.7061

Moulieswaran, N., & Prasantha Kumar, N. S. (2023). Investigating ESL learners' perception and problems towards artificial intelligence (AI)-assisted English language learning and teaching. *World Journal of English Language*, *13*(5), 290-298. https://doi.org/10.5430/wjel.v13n5p290

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. U.S. Department of Health, Education, and Welfare. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Pegrum, M. (2014). *Mobile learning: Languages, literacies and cultures*. Palgrave Macmillan. https://doi.org/10.1057/9781137309815

Seedhouse, P. (2009). The interactional architecture of the language classroom. *Bellaterra Journal of Teaching & Learning Language & Literature, 1*(1), 1–13. https://doi.org/10.5565/rev/jtl3.32

Su, J., Ng, D. T. K., & Chu, S. K. W. (2023). Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence, 4,* Article 100124. https://doi.org/10.1016/j.caeai.2023.100124

Tseng, T., Murai, Y., Freed, N., Gelosi, D., Ta, T. D., & Kawahara, Y. (2021). PlushPal: Storytelling with interactive plush toys and machine learning. *Proceedings of the 20th International Conference on Interaction Design and Children* (pp. 236–245). Association for Computing Machinery. https://doi.org/10.1145/3459990.3460694

Williams, R., Park, H. W., Oh, L., & Breazeal, C. (2019). PopBots: Designing an artificial intelligence curriculum for early childhood education. *Proceedings of the 33rd AAAI*

Conference on Artificial Intelligence, *33*(1), 9729–9736. Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v33i01.33019729

Woo, J. H., & Choi, H. (2021). *Systematic review for AI-based language learning tools.* arXiv. https://doi.org/10.48550/arXiv.2111.04455

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3), Article 212. https://doi.org/10.3390/languages8030212

Appendix A

Table A1

Transcription key (adapted from Jefferson, 2004)

text?	intonation that clearly indicates a question
/	rise in intonation
\	drop in intonation
(.)	very brief pause (less than 0.5 seconds)
>text<	faster than normal utterance
[text]	overlapping utterances
text:	syllable is slightly elongated
TEXT	louder than the rest of the words
((text))	transcriber notes about actions
text	spoken in Catalan
<u>text</u>	spoken in Spanish
text	translation of Catalan or Spanish

About the Authors

Maria Mont is an award-winning English teacher, translator, and interpreter with nearly 20 years of experience in education. She holds a Master's degree in Education and ICT, specializing in ICT tools, Artificial Intelligence, and telecollaborative projects to enhance language learning. Maria has worked as an educational advisor for the Catalan Government and as an adjunct lecturer at Universitat Autònoma de Barcelona. Currently pursuing a Ph.D. at UAB, her research focuses on AI in education, particularly language teaching. A published author and keynote speaker, she has presented at international conferences, including ICSS XVI in Paris and TSLL 2024 in Iowa.

Dr. Melinda Dooly is a Serra Húnter Full Professor in the Department of Language & Literature Education and Social Science Education at the Universitat Autònoma de Barcelona. Her research principally focuses on technology-enhanced project-based language learning, intercultural communication, and the development of 21st century competences in teacher education. She has published extensively in international journals and is the author and editor of numerous books and chapters in these areas. She is currently co-editor of the book series Telecollaborative Learning and Virtual Exchange in Education (formerly Telecollaboration in Education), published by Peter Lang. In recognition of her research trajectory, she has recently been awarded a five-year ICREA Acadèmia fellowship by AGAUR (2025-2030). She is a frequent keynote speaker and invited lecturer at academic events worldwide and currently serves as a member of the Research Council for Education at the Spanish National Agency for Investigation. https://www.melindadooly.com/

Onyemowo Ejeme is a teacher, humanitarian, and educational consultant with over a decade of experience in curriculum development, teacher training, and school management. As the founder of Oakbridge Foundation, she provides free education for indigent learners, promoting inclusion, diversity, and equal access to quality education. For further discussion and collaboration, connect on http://bit.ly/4hmh12V or visit www.oakbridgefoundation.org

Chapter 7

Unlocking the Potential of Generative AI: Enhancing Spanish Language Learning through Humor Translation Strategies

Cristina Pardo-Ballester
Iowa State University

This study investigated how generative AI (GenAI) can support students in translating humor from Spanish to English, a task that is notoriously complex due to cultural nuance, wordplay, and the need to preserve the intended humorous effect. The problem addressed is that humor translation requires creative, context-sensitive adaptation, a process often time-consuming and challenging for students with limited experience. While GenAI offers linguistic and semantic suggestions, it often falls short in accurately capturing subtle cultural references or comic intent without precise user guidance. Many students initially struggled with crafting effective prompts and critically interpreting GenAI outputs, which sometimes resulted in humorless or inaccurate translations. Through contrastive analysis and the General Theory of Verbal Humor (Attardo, 2020), the study demonstrated that successful translations preserve script opposition and logical mechanisms while adapting language and cultural context to resonate with the target audience. GenAI played a central role as a creative scaffold, actively assisting students in exploring multiple humorous renderings, guided by dynamic equivalence (Nida, 1964) and communicative function (Baer, 2017). Findings suggest that when students develop sophisticated prompt engineering strategies and critical evaluation skills, GenAI becomes an indispensable and meaningful support in the nuanced process of humor translation.

Cite as: Pardo-Ballester, C. (2025). Unlocking the potential of generative AI: Enhancing Spanish language learning through humor translation strategies. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 122-144). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.07
©2025 Pardo-Ballester. Published under a CC BY license.

Introduction

In December 2022, while listening to the NPR Up First podcast, I learned about ChatGPT, a generative AI (GenAI) chatbot capable of content creation, knowledge discovery, and conversational AI. Intrigued, I registered immediately and began testing its capabilities for question-answering, translation assistance, and Spanish task creation. ChatGPT, developed by OpenAI, is a large language model (LLM) trained on multilingual datasets to generate coherent and contextually relevant responses (OpenAI, 2023). Its functions extend beyond simple translation to include content creation, knowledge discovery, and conversational AI. Scholars like Muñoz-Basols et al. (2023) suggest that AI-powered translation tools, when paired with pedagogical guidance, can enhance critical thinking and intercultural competence in language learning. This aligns with Nacenta's emphasis that AI's creative potential truly emerges through user interaction, positioning GenAI as a 'co-creator' offering students multiple translation hypotheses to critically analyze and refine (CCCB, 2023). However, these tools face notable limitations. Floridi and Chiriatti (2020) argue that although LLMs can generate text that appears coherent and human-like, they rely on statistical prediction rather than genuine semantic understanding, which impairs their ability to interpret pragmatic context, cultural nuance, or humor. Su and Goslar (2024) further emphasize this limitation, showing that LLMs struggle with pragmatic competence due to their inability to infer subtle contextual cues without explicit semantic input. This shortcoming is especially apparent in humor translation, a task that demands creative reason, cultural adaptation, and script manipulation. Popova and Dadić (2023) note that even human translators face challenges in conveying humor, which often relies on cultural references and puns. Current GenAI tools lack dedicated mechanisms for handling such complexities, and successful translations of humor by machines typically occur by coincidence rather than true comprehension, underscoring the rudimentary state of GenAI's comedic capabilities.

As a Spanish instructor, I have been integrating machine translation (MT) in an introductory Spanish-English translation course for years (Pardo-Ballester, 2021, 2022), especially for expressive texts such as short stories and jokes. Humor, however, poses unique challenges due to its dependence on linguistic ambiguity, incongruency, and audience expectations (Baer, 2017; Vandaele, 2010). Cultural context significantly affects how humor is understood and conveyed, making effective translation difficult, particularly for students still developing linguistic and cultural fluency.

Despite these limitations, the complexity of humor translation presents a significant challenge that traditional pedagogical approaches, constrained by time for feedback and redrafting, often struggle to address. This is where GenAI offers a compelling potential to bridge the gap. By generating draft translations and suggesting alternative phrasings in real-time, GenAI can significantly ease the burden of early-stage drafting, thereby creating more space for students' creativity and experimentation within time-limited classroom settings. Its capacity to provide multiple output options encourages students to manipulate scripts and adapt cultural references more freely, provided they are guided in how to prompt effectively.

Baer (2017) highlights that students must consider the communicative function and target audience when translating, often needing to justify their translation choices. Nida's (1964)

principle of dynamic equivalence supports adaptations that evoke a similar response in the target audience and cultural resonance must be preserved.

Several traditional methods are employed when translating humor, each with its own approach and focus. However, Attardo (2002) observes a significant gap in translation theories, particularly in relation to humor. He notes that these theories lack a systematic way to measure semantic similarity between source and target texts. This deficiency is particularly evident in humor translation theories, which have yet to develop their own metrics for assessing this similarity. That is why Attardo (2002) suggests that the General Theory of Verbal Humor (GTVH), developed by Attardo and Raskin (1991), offers a solution by including a metric for measuring similarity in humorous texts across languages. This theory is valuable for humor translation, where preserving the humorous effect is crucial.

Incongruency, the mismatch between expectations and reality, plays a crucial role in humor, as it often creates the surprise and absurdity that make jokes funny. Vandaele (2010) explains the incongruity approach with this joke:

> The solution to the incongruity is often cognitively farfetched (yet locally relevant in a given discourse): Is the doctor home?" the patient asked in his bronchial whisper. "No," the doctor's young and pretty wife whispered in reply. "Come on in." (Raskin, 1985, p. 100). In the context of a doctor's visit (the set-up), the behavior of the wife is farfetched yet understandable via a radical reframing of the action situation. (p. 149)

According to Attardo (2002), the GTVH analyzes humor using six hierarchical knowledge resources (KRs) "to determine how different a source text is from its translation, humor-wise" (Attardo, 2020, p. 245.) I will illustrate the KRs with the joke of the doctor presented above:

1. Script Opposition (SO): Humor arises when a joke juxtaposes two conflicting scripts, such as a doctor visit and a sexual encounter; or a doctor and a lover. The phrase "Come on in" serves as the punchline, shifting the script from medical to romantic.
2. Logical Mechanism (LM): The joke connects the opposing scripts through incongruity resolution, making sense in a playful way. The LM is the misunderstanding. The patient's whisper suggests a medical condition, while the wife's whisper and invitation suggest a romantic encounter.
3. Situation (SI): The setting and context of the joke involves a patient visiting a doctor's home and interacting with the doctor's wife. The bronchial whisper adds to the context, making the initial medical script plausible.
4. Target (TA): The target subject of the joke (a person, group, or institution) could be seen as the doctor's wife, who is portrayed in a flirtatious manner, or the patient, who is unwittingly caught in an awkward situation.
5. Narrative Strategy (NS): The joke uses a dialogue format with a setup (the patient's question) and a punchline (the wife's response).

6. Language (LA): The use of "bronchial whisper" and "whispered in reply" are key linguistic elements that set up the dual scripts and contribute to the humor. The whispering adds to the humor by creating a sense of secrecy and intimacy.

While traditional humor translation theory has not explicitly accounted for AI, recent developments in GenAI, including LLMs and MT systems, offer promising methods that align well with established translation principles such as Nida's (1964) dynamic equivalence and Baer's (2017) cultural adaptation. GenAI enhances creative processes by providing alternative interpretations that students can evaluate and refine. As López de Mántaras (2023) notes, AI lacks fundamental human faculties such as common-sense reasoning and cultural intuition, making human oversight indispensable. López de Mántaras (2025) reiterates that AI is not truly intelligent or capable of genuine understanding; its utility lies in its ability to interact with human users who guide, correct, and interpret its suggestions.

One of the main challenges in humor translation is preserving semantic accuracy, cultural relevance, and the humorous effect. These tasks typically require time, expertise, and multiple rounds of revision. In educational contexts constrained by limited time for feedback and redrafting, these demands can overwhelm both students and instructors. GenAI addresses this issue by offering real-time assistance: generating draft translations, suggesting alternative phrasings, and helping students visualize how wordplay or cultural references might be received by different audiences.

GenAI supports creativity and experimentation by alleviating the demands of initial drafting, particularly within the time limitations inherent to online instruction. Its capacity to provide multiple output options encourages students to manipulate scripts and adapt cultural references more freely, provided they are guided in how to prompt effectively. Rather than automating the process of humor translation, GenAI supports key phases of the translation process: comprehension, experimentation, and revision.

The theoretical perspectives of Baer (2017), Nida (1964), and Attardo (2002, 2020) offer a foundation framework for evaluating GenAI's pedagogical value in this context. Baer's (2017) emphasis on communicative function and audience positioning supports the use of GenAI as a tool for generating culturally relevant translation variants that students can critically assess. While GenAI does not independently apply translation theories, Nida's (1964) distinction between formal and dynamic equivalence provides a lens through which students can evaluate AI-generated outputs, distinguishing between literal renderings and those that aim to preserve rhetorical or emotional impact. In this way, GenAI serves not as an autonomous translator but as a resource for students to explore multiple strategies and assess functional effectiveness in humor translation. Attardo's GTVH adds a concrete framework for evaluating whether GenAI outputs maintain essential components of humor, such as script opposition, logical mechanisms, and narrative strategy. Together, these theories allow students to approach GenAI-generated translations not as final solutions but as exploratory possibilities to be assessed functionally and rhetorically.

This study addresses a gap in translation pedagogy by examining how GenAI, when thoughtfully integrated, can support the development of humor-sensitive translation skills while simultaneously easing the demands of initial creative production. Specifically, it investigates the

challenges and opportunities of incorporating GenAI into the Spanish-language classroom, with a focus on humor translation. This study is guided by the following question:

> What challenges do students face in humor translation, and how do these affect their ability to convey cultural nuances and dynamic equivalence?

While humor translation has long been recognized as a complex task requiring cultural adaptation and rhetorical sensitivity, little research has explored how GenAI tools can scaffold this process in pedagogical settings. This study responds to that need by examining how GenAI can serve as a creative, dialogic partner, helping students not only produce translations but also reflect critically on the pragmatic and intercultural dimensions of humor.

Methods

This study focuses on the role of GenAI in supporting humor translation in an intermediate-level Spanish-English online translation course. The goal was to understand how GenAI tools could support students in interpreting and adapting jokes for an American audience. To ensure a systematic and pedagogically sound approach, this study used a design-based research framework (Pardo-Ballester, 2021, 2022), incorporating reflections, feedback, and iteration.

The study incorporated multiple GenAI tools: ChatGPT, Google Gemini, and Microsoft Copilot. These were chosen strategically to explore how different GenAI systems handle linguistic ambiguity, cultural context, and humor mechanisms. Unlike traditional MT systems like DeepL (DL) and Google Translate (GT), which were used for initial comparisons (see Step 1 under 'Procedures'), GenAI tools provide dialogic and interactive capabilities. These features allowed students to engage with the tools by requesting clarifications, exploring cultural nuances, and experimenting with humor-enhancing phrasing, thereby fostering a more dynamic and learner-centered translation process.

The use of multiple GenAI platforms facilitated a comparative understanding of how various models process humor and supported student agency in evaluating and selecting the most appropriate output. Rather than serving as a replacement for human interpretation, GenAI functioned as a co-creative partner. It enabled students to explore alternative phrasing, test hypotheses, and iteratively revise their work. These affordances, including real-time prompting, semantic explanations, and cultural sensitivity, position GenAI as a pedagogically valuable tool for humor translation, exceeding the capabilities of conventional MT systems.

Although quantitative evaluation was not the primary aim of the study, GenAI's impact was examined qualitatively through triangulated data sources: student reflections, final translations, and the prompt-output chains submitted throughout the assignment. This approach provided an in-depth understanding of how students leveraged GenAI to navigate challenges in humor translation, such as identifying linguistic ambiguities, assessing cultural equivalents, and refining rhetorical and stylistic choices.

Participants

Seventeen undergraduate students (14 women, 3 men) enrolled in an online Spanish – English translation course at Iowa State University participated in the study during the Fall 2024 term. All students were at an intermediate-mid proficiency level in Spanish and had prior exposure to MT but no formal training in GenAI technologies.

Procedures

Students engaged in two humor translation tasks, each based on a Spanish language joke from a mid-20th-century text and paired with an image. Tasks were structured across six steps (see Appendix A):

A. Reading and understanding (Steps 1-3): Students first reviewed the original Spanish joke and observed literal translations produced by GT and DL. These were presented alongside back-translations to reveal how MT systems distort humor when context and pragmatic are ignored.
B. AI interaction (Step 4): Students then generated their own prompts and interacted independently with ChatGPT, Gemini, and Copilot. They were explicitly instructed to explore not just direct translations, but also cultural equivalents, tone adaptations, and explanatory clarifications. Examples show that students often asked the AI to explain double meanings, translate cultural references like "¡Ave María Purísima!", and improve comedic impact.
C. Comparison and creative translation (Step 5): Students compared GenAI outputs with MT outputs to assess which options best captured the humorous effect and cultural framing of the original. They then submitted their final, creative English translation.
D. Reflection (Step 6): Students reflected in Spanish on their process, describing the problems they faced, how they used AI to resolve them, and the translation methods they adopted. They also considered how the accompanying cartoon influenced their decisions.

In this study, GenAI was not used to produce final translations but rather functioned as a comprehension and ideation tool. Students were given full agency to craft their own prompts when interacting with ChatGPT, Gemini, and Copilot, rather than working from predefined queries. This approach emphasized prompt literacy as a key learning outcome, helping students understand how their input shaped GenAI's humor, tone, and cultural appropriateness. This design decision encouraged students' ownership of the translation process and introduced them to prompt engineering, helping them understand how different prompts yield varied levels of cultural and linguistic appropriateness. The instructional materials explicitly emphasized that well-constructed prompts could elicit more refined, humorous, and context-sensitive outputs, transforming GenAI into a tool for exploration rather than automation.

Students assessed GenAI outputs to determine which suggestions preserved the original humor and made conscious decisions about what to select, reject, or revise. These outputs were shared with the instructor along with final translations and written reflections.

Results

Challenges in Humor Translation: Impact on Students' Ability to Convey Cultural Nuances and Achieve Dynamic Equivalence

A contrastive analysis of the source texts and their translated texts served to categorize responses based on participants' successful or unsuccessful adaptation of linguistic and cultural nuances preserving the joke's humor. In addition, this analysis served to illustrate the similarity of jokes across languages applying the GTVH.

Table 7.1 presents the first cultural joke (referred to as 'the optimist joke') that participants translated, which relies on the Spanish tradition of celebrating April Fool's Day on December 28th (the joke is reproduced in Appendix A, Translation Activity 1). A successful translation required recognizing this cultural difference and adapting the date for an American audience, as well as understanding the implicit meaning of "*con una*" [with one] as 'with another woman.' Linguistically, 82.35% of students successfully conveyed the implicit meaning of "*una*" as [woman] or [chick] in their translations, often directly aided by GenAI which provided clarifications or options like 'with another woman,' 'with a different woman,' 'with a woman,' or 'with some woman.' For instance, ChatGPT outputs for Joke 1 consistently translated "*con una*" as 'with someone' or 'with another woman,' and Gemini outputs similarly provided 'with a woman,' or 'with another woman.' This indicates GenAI's ability to assist students in discerning implicit semantic meaning.

Table 7.1

Cultural Joke Translation Using a Communicative Approach

Joke 1 Source text	Joke 1 Target text
Una optimista *- ¡Me río de lo tontísimo que es mi marido! Figúrese que acabo de verlo en un taxi con una.* *- ¡No sería él!* *- Sí, hombre, sí.* *¡Es que él se figura que yo no sé que estamos a 28 de diciembre!*	An optimist - I laugh at how ridiculously silly my husband is! Just imagine, I just saw him in a taxi with another woman! - It could not be him! - Yes, it was him. He thinks I don't know that it's April First!

While 64.70% of students did not submit their assignments on time to receive instructor feedback, and of the 35.29% who did, only 17.64% incorporated the instructor's comments to improve their translations. This small percentage of students who successfully integrated the feedback demonstrated an enhanced ability to leverage GenAI through more sophisticated prompting and critical assessment, ultimately leading to translations that more effectively captured the original humor and cultural nuances for an American audience.

Culturally, 58.82% of participants, with the assistance of GenAI, recognized the cultural difference of December 28th being akin to April Fool's Day in Hispanic cultures. GenAI tools provided explicit explanations for this cultural nuance in response to student prompts. For

example, student prompts asking 'what does December 28th mean?' (student 1) or providing context like 'knowing that the reference December 28 means like an April fools day' (student 2) directly led to AI outputs explaining "*Día de los Santos Inocentes*" [the Day of Holy Innocents] and its equivalence to April Fool's Day. However, these participants still struggled to adapt the joke effectively for an American audience, leading to translations that lacked humor. This difficulty manifested as retaining 'December 28,' omitting the date entirely, or adding explanatory information such as 'the Day of Holy Innocents,' which is not widely known in American culture, ultimately rendering the jokes nonsensical for an American audience.

Culturally, only 41.18% of participants successfully adapted the date in their final translations by using 'April Fool's Day' or 'April 1st.' This success was often directly linked to their ability to leverage GenAI through effective prompting. For instance, student 1's final translation successfully used 'April First,' a result influenced by ChatGPT's explanation of December 28th's meaning. Student 8 similarly produced a translation using 'April Fool's Day' after actively prompting ChatGPT for the significance of the date and asking it to incorporate that information. These findings indicate that students' success in achieving dynamic equivalence (Nida, 1964) required not only a thorough understanding of the source material, but also a creative approach to adapting the humor for the target language, often facilitated by iterative prompting with GenAI. This finding indicates that students tend to translate jokes in a manner that aligns more closely with the hierarchical structure proposed by the GTVH (see Table 7.2). This example illustrates how GenAI's output aligns with the GTVH concept of script opposition.

Table 7.2

The Six Knowledge Resources of the GTVH for the Optimist Joke

KRs	Joke 1: An optimist
SO	Infidelity vs. Innocent Prank, reality vs. perception
LM	Misunderstanding/Prank
SI	Wife sees husband in a taxi with another woman on December 28th/April Fool's Day
TA	Husband (portrayed as foolish)
NS	Dialogue with setup and punchline
LA	Key phrases: "*tontísimo*," [ridiculously silly] "*con una,*" [with a woman] "él se figura...," [he thinks], "*28 de diciembre*" [April First]

Participants with unsuccessful translations due to difficulty in finding creative equivalents struggled transferring all the KRs, when the translation was too explicit, or the joke was translated with significant alterations. For example, participant 11 introduced a religious metaphor, and the lack of clear reference to the cultural context of December 28th disrupted the original joke's humor and coherence. Consequently, the reference to 'waltz his way through the doors of God' introduced a different script that didn't align with the original joke's context:

"The Optimist:

- I'm dying of laughter at how dumb my husband is!
  I just saw him get inside a taxi with another woman.
- That's not him!
- My friend, do you think I was born yesterday?
  He just thinks he can waltz his way through the doors of God just by being a pig!"

The translation above differs at the level of SO which is the higher level in the GTVH's proposed hierarchy, leading to greater dissimilarity between the source and target text. This result is in line with Attardo and Raskin's (1991) study that predicted that changes in lower-level KRs, such as LA, would result in a higher degree of perceived similarity than changes in higher-level KRs, such as SO.

Table 7.2 above shows the six KRs of the GTVH for the optimist joke (see Appendix A, Translation Task 1). As Attardo (2020) states the KRs prove that the similarity exists at a deeper conceptual level than mere surface wording. This analysis shows a similarity in how humor is constructed and appreciated across languages. It helps understanding the effectiveness of various humor techniques and their impact on the American audience. In this analysis, the joke involves two OSs: Infidelity and innocent prank, or we can also see it as reality and perception. The initial script suggests the husband might be unfaithful, but the punchline reveals it is a harmless prank related to December 28th, which is the Spanish equivalent of April Fool's Day. The LM is the misunderstanding or mistaken belief. The wife pretends to believe her husband is unfaithful, but the punchline reveals she knows it is a prank day, creating humor through the twist. The SI involves a wife talking about her husband being seen in a taxi with another woman. The context of December 28th is crucial as it sets up the prank. The TA of the joke is the husband, who is portrayed as foolish for thinking his wife wouldn't realize it is a prank day. The wife also humorously highlights her own awareness and cleverness. The joke uses a dialogue format with a setup (the wife's observation and friend's disbelief) and a punchline (the wife's revelation about December 28th). The dialogue builds up the expectation of infidelity before delivering the twist. The LA includes phrases like "*¡Me río de lo tontísimo que es mi marido!*" [I laugh at how ridiculously silly my husband is!] and "*él se figura que yo no sé*," [He thinks I don't now] which set up the humor. The use of "*28 de diciembre*" [April First] is key to the punchline, as it relies on cultural knowledge of the prank day.

Table 7.3 presents the second linguistic joke, "El auto de línea," [line car] (see Appendix A, Translation Task 2). This joke relies on a pun between the similar-sounding Spanish words "*baca*" [roof rack] and "*vaca*" [cow]. The core humor lies in the unexpected shift from the 'roof rack' script to the 'cow' script, making it challenging to find direct English equivalents. The title is the first difficulty to translate because students have a cartoon featured with a caption. "El auto de línea" which literally is [line car] looks more like a small bus (see Appendix A for the cartoon). All participants, except for one, forgot to translate the title. This joke utilizes a pun, playing on the similar sounding Spanish words. The difficulty in translation arises because this opposition is tied to the specific sounds of the Spanish language and the lack of direct equivalents in English. When we hear the sentence "*subir a la baca*," and paying attention only to the sound it can mean [get on the roof rack/cow] or [load the cow]. An effective translation needs to find a comparable English pun or a wordplay that maintains the humor which can be very complex to accomplish. In addition, it is important to recognize other expressions such as

"*Ave María Purísima*" to express surprise which literally means [Hail Mary Most Pure] and we can translate it for the more modern expression [My Goodness!] "*Gracias a Dios*" [Thank God] is another expression that illustrates that religion was an integral part of the daily life in Spain and in the 1950s, reflecting the deep-rooted cultural and societal norms of that era.

The translation provided in Table 7.3 attempts a wordplay with 'moo-ve' using an onomatopoeia. However, it does not fully replicate the original pun and differs at the original SO, which is the higher-level KRs that Attardo (2002, 2020) emphasizes must be preserved. 76.47% of students did not submit their translation before the first due date to receive feedback. Only 23.52% of the participants interacted with their instructor, received feedback and improved their target translation.

Table 7.3

Adapting a Linguistic Joke to Create an Effective Equivalent Effect in Translation

Joke 2 Source text	Joke 2 Target text
El auto de línea	Community bus
- Como haber sitio, sí lo hay; pero va usted a tener que subir a la baca.	- Well, there is room, but you'll have to travel on the cow.
- ¡Ave María Purísima! Gracias a Dios, la vaca se queda.	- Oh, Goodness, I did not expect that! I guess that's one way to moo-ve!

Recognizing the pun was more challenging than expected; only 58.82% of participants identified the pun "*baca vs. vaca*" [roof rack vs. cow]. Within this group of participants who successfully identified the pun, GenAI played a crucial role in helping students grasp this nuance and explore creative solutions. AI outputs frequently explained the pun directly, for example, Copilot stated: 'It's a play on words where "baca" (roof rack) is confused with "vaca" (cow).' Gemini also provided explanations and suggested literal translations with the pun clarified. This direct explanation by GenAI allowed students to conceptualize the joke beyond a literal translation. This group successfully adapted the linguistic terms, either by using the context of the cartoon, finding another pun, or by using an explicit translation (participant #16) demonstrating a good understanding of the puns:

> "Man: If you want to get in, you'll have to ride on the roof rack.
> (The woman turns and looks back at the cow)
> Man: No, you!
> Woman: Oh, thank God! I'm going on the roof rack and not the cow!" (participant #16)

The successful translation demonstrated similarity in the opposing scripts within the GTVH framework (see Table 7.4). For example, participant # 5, after prompting for a similar English joke where words sound alike but mean different things, produced a translation using the pun 'bear' and 'bare,' highlighting the homophony:

> "- Do you have any room for one more?
> - Sure, but you'll need to 'bare' with us a bit.

- Oh no! I was hoping for a space, not to get bare."

Participant #14 consistently opposed scripts and translated them with creativity, employing clever wordplay. For instance, the transformation of 'cargo rack' to 'cow-go-rack' introduces a humorous and unexpected twist. The humor stems from the surprise and absurdity of the second script, which contrasts sharply with the more mundane nature of the first script:

"- Do you have any room for one more?
- We have space, but only on the cargo rack.
- A cow-go-rack? Holy Mary! Thank God, that's perfect for my cow!"

Participant #14 demonstrated creativity after GenAI's explanation of the pun. This was a result of interactive prompting where GenAI suggested wordplay and explanations that the student then built upon, such as Gemini's suggestion of 'cow-abunga!' as a playful reference.

Participant #6, through repeating prompting, also explored homophones like 'ox/box' to adapt the joke. Similarly, participant #1 found GenAI helpful in suggesting 'moo-ve' as a pun. These examples highlight how GenAI acted as a co-creator, providing multiple translation hypotheses that students could analyze and refine.

In contrast, 41.18% of students were not able to adapt the joke, primarily due to an inability to replicate the pun, leading to literal translation. This confirms the inherent difficulty of humor translation when cultural or linguistic specificities are not successfully navigated, even with GenAI assistance. These results underscore that while GenAI can generate options, human critical assessment and creative intervention remain indispensable.

Table 7.4 shows the six KRs of the GTVH for the bus joke. With this analysis we can see how each element contributes to the overall humor. The SO involves the opposition of the pun between "*baca*" (available space on the bus) and "*vaca*" (misunderstanding with loading the cow). The adapted translation also involves a SO between the pun "bear" (to be patient) and "bare" (to be unclothed). The humor arises from the unexpected shift in meaning. These puns are central to the joke, and the dialogue format helps deliver the punchline effectively. The humor arises from the unexpected twist where the listener confuses the words leading to an absurd and funny conclusion. The LM is the pun based on the phonetic similarity to create a humorous misunderstanding which is resolved in the punchline, making the joke effective. The SI involves a conversation where a woman is asking for space on a bus and the possibility of having to sit on the roof rack. With the translation, the SI involves a woman asking for space and she is told to be patient, while the absurdity of the misunderstanding adds to the humor. The TA is a woman who becomes the butt of the joke due to their misinterpretation. The NS is a dialogue structure that builds up the expectation and delivers the punchline effectively, enhancing the humor impact. The joke is presented as a conversation between two people, which helps set up (the statement about available space) and a punchline (the exclamation about the cow or to be naked). The LA used is crucial for the pun. The specific choice of words ("*baca* vs. *vaca*" and 'bear vs. bare') which sound similar but have different meanings.

Table 7.4

The Six Knowledge Resources of the GTVH for the Bus Joke

KR	Joke 2: A community bus
SO	*Available space* vs. misunderstanding; to be patient vs. to be unclothed
LM	A pun ("*baca*" vs "*vaca*"); or ('bear' and 'bare') to create a humorous misunderstanding.
SI	A conversation about bus space with the bus driver.
TA	The butt of the joke is a woman.
NS	Dialogue with setup and punchline
LA	Key words ("baca" and "vaca"/ 'bear with us' and 'bare') and cultural exclamations

The complexities of language, culture, and humor require translators to be adaptable and creative, employing various strategies to ensure the translated joke effectively conveys the intended humor and meaning. Moreover, Attardo (2002) highlights "let your translation differ at the lowest level necessary for your pragmatic purposes" (p.183), but when translating to ensure its humorous function in the new context reaching similarity at the highest level is a complex task.

Participants found that literal translations often failed to capture humor, highlighting the issue of literal versus functional translation. Some students moved away from literal translations to prioritize the humorous effect. MTs were also unsuccessful at rendering humor. Additionally, participants who struggled to understand the deep meaning of both jokes demonstrated the difficulty of transferring two opposing scripts, which is key to understanding humor within the GTVH.

GenAI's Role and Prompt Engineering

One of the significant challenges observed in these translation activities was crafting an appropriate prompt or engaging sufficiently with GenAI to grasp the deep meaning of the joke. Participants who wrote effective prompts were able to understand the significance of the Spanish date December 28 and the pun of "*baca*" [roof rack] vs. "*vaca*" [cow]. Only 35% of the participants showed an adequate understanding of prompt usage. These participants provided detailed instructions to the AI, explicitly considering cultural context, wordplay, and the need for adaptation to the American audience. For instance, Participant # 4's prompt for Joke 1, 'Translate the following Spanish joke considering the historical, cultural, and sociological context of the Spanish language to make the joke both funny and comprehensible in English,' demonstrates a high level of prompt literacy. Participant # 5 for joke 1 also specifically requested adjustment for 'U.S. humor, language and dates.' Similarly, participant # 11 for joke 2 included extensive context in her prompt, identifying as a translation student and asking for cultural, historical, sociological, and deep meaning for a U.S. audience. These well-constructed prompts often elicited more refined, humorous, and context-sensitive outputs from GenAI.

Consistent and specific instructor feedback reflected a clear pedagogical effort to support students in developing more sophisticated prompt engineering strategies. This guidance directly addressed a key limitation of GenAI: its inability to interpret pragmatic context or cultural nuance without explicit semantic input. The students' improved ability to craft elaborated prompts can therefore be attributed to the iterative learning process and ongoing instructional support provided throughout the course.

In contrast, 65% of participants showed a limited understanding of prompt usage. They often used vague or general prompts that did not provide enough information to the AI, such as 'translate this joke into English' (participant #1, joke 2) or simply pasting the text (participant #15, joke 1). These prompts failed to consider cultural context or wordplay, resulting in literal translations rather than ones that conveyed the humor of the joke.

The study revealed that while participants recognized the potential of AI tools in providing information and insights, they also identified limitations: GenAI's lack of inherent cultural sensitivity, its tendency towards literal and awkward translations, its inclusion of unnecessary details, and the absence of essential human elements like genuine understanding and creativity. However, when students actively engaged in a co-creative process with GenAI, exploring multiple hypotheses and testing different prompts, AI proved to be a powerful tool for developing their critical and creative translation skills.

Conclusion

This study provides valuable insights into the challenges and opportunities of integrating GenAI in humor translation for Spanish language learning. It underscores that while GenAI tools can offer significant assistance in generating linguistic and semantic suggestions, their full potential in humor translation is unlocked only through effective prompt engineering and critical human evaluation. The findings highlight the importance of teacher feedback in guiding students, and perhaps more significantly, the irreplaceable role of human creativity and cultural insight in achieving truly effective and humorous translations. As Lee (2024) suggests, the evolving landscape of AI requires translators to reevaluate their role, focusing on cross-cultural communication projects, post-editing, and consulting, leveraging AI's capacity for contextual understanding and handling complex linguistic situations. For instructors, this means moving beyond traditional criteria of linguistic accuracy and translation quality to include the cultivation of prompt engineering skills, critical-creative thinking, and a deeper understanding of intercultural communication.

To further advance this field, future research should explore the development of pedagogical strategies and resources specifically designed to enhance student understanding of prompt usage and AI capabilities for nuanced tasks like humor translation. This includes designing AI tools that are more sensitive to cultural nuances and humor styles and examining long-term impact of AI integration on student translation skills, creativity, and cultural sensitivity.

By addressing these areas, we can further unlock the transformative potential of GenAI in language education, fostering not just linguistic proficiency but also deep intercultural

understanding and creative problem-solving abilities. This integration of GenAI tools presents a unique opportunity to enhance students' skills, while simultaneously highlighting the ongoing need for human insight, creativity, and cultural sensitivity in every step of the translation process.

References

Attardo, S. (2002). Translation and humour: An approach based on the General Theory of Verbal Humour (GTVH). *The Translator*, *8*(2), 173-194. https://doi.org/10.1080/13556509.2002.10799131

Attardo, S. (2020). *The linguistics of humor: An introduction*. Oxford University Press. https://doi.org/10.1093/oso/9780198791270.001.0001

Attardo, S., & Raskin, V. (1991). Script theory revis(it)ed: joke similarity and joke representation model. *HUMOR*, *4*(3-4), 293-348. https://doi.org/10.1515/humr.1991.4.3-4.293

Baer, B. J. (2017). Teaching translation through text types. In L. Venuti (Ed.*), Teaching translation: Programs, courses, pedagogy* (pp. 63-70). Routledge. https://doi.org/10.4324/9781315623139-10

CCCB. (2023, November 7). *A morning with Lluís Nacenta: AI and creativity* [Video]. CCCB. https://www.cccb.org/en/multimedia/videos/a-morning-with-lluis-nacenta/243835

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds & Machines*, *30*, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Lee, T. K. (2024). Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review*, *15*(6), 2351–2372. https://doi.org/10.1515/applirev-2023-0122

López de Mántaras, R. (2023, February 14). *Inteligencia Artificial: Grandes Retos* [Video]. YouTube. https://www.youtube.com/watch?v=1EA7ZejUcJQ

López de Mántaras, R. (2025, February 17). *¿Es realmente inteligente la Inteligencia Artificial?* [Video]. YouTube. https://www.youtube.com/watch?v=CA8LdZpG5l0

Muñoz-Basols, J., Neville, C., Lafford, B. A., & Godev, C. (2023). Potentialities of applied translation for language learning in the era of artificial intelligence. *Hispania 106*(2), 171-194. https://doi.org/10.1353/hpn.2023.a899427

Nida, E. (1964). *Towards a science of translating*. Brill.

OpenAI (2023). *GPT-4 technical report*. https://openai.com/research/gpt-4

Pardo-Ballester, C. (2021). “Design-based research”: Para el aprendizaje de una lengua en un curso de traducción. *E-AESLA*, *7*, 1-15. https://perma.cc/NA8L-HLS9

Pardo-Ballester, C. (2022). A case study: A technological insight on teaching translation. *Foreign Language Annals*, *55*(3), 894–913. https://doi.org/10.1111/flan.12640

Popova, O., & Dadić, P. (2023). Does AI have a sense of humor? CLEF 2023 JOKER Tasks 1, 2 and 3: Using BLOOM, GPT, simpleT5, and more for pun detection, location, interpretation and translation. *CEUR Proceedings*, *3497*, 1–21. https://perma.cc/86PJ-AMNQ

Raskin, V. (1985). *Semantic mechanisms of humor.* Springer Dordrecht. https://doi.org/10.1007/978-94-009-6472-3

Su, D., & Goslar, K. (2024). Artificial intelligence shows human-like lens capability in subjectivity: Evaluating Chinese pragmatic capability of GPT-4 via Microsoft Copilot with direct prompting. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 175–201). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.11

Vandaele, J. (2010). Humor in translation. In Y. Gambier, & L. Van Doorslaer (Eds.)*, Handbook of translation studies* (pp. 147-152). John Benjamins. https://doi.org/10.1075/hts.1.hum1

Appendix A

Translation Activity 1: Una Optimista.

Context:

The source text, "*Una optimista,"* is from a book published in the 1950s in Spain. Google Translate (GT) and DeepL Translator (DT) were used to translate the source text from Spanish into English. The English results using machine translation (MT) are found in Step 2.1(GT) and Step 3.1 (DT). Both are literal, word-for-word translations, where readers are aware they are reading a translation. MT relies only on surface-level features, ignoring nuanced word meanings and the complex context of humor. MT was used to back-translate the English texts (Steps 2.1 and 3.1), resulting in new Spanish versions (Steps 2.2 & 3.2).

Note: Understanding historical, cultural, and sociological differences between languages is essential for effective translation. Use all available techniques, such as monolingual dictionaries, artificial intelligence models, or consultation with native speakers, to identify double meanings within specific contexts that reflect cultural values. Once you understand the text beyond the surface level, you can produce a translation that recreates the humor, ideally preserving the wordplay.

You have three or more attempts for this exercise. The due date is Sunday. To receive instructor feedback, submit your first attempt by Thursday (midnight). Submissions after this deadline will not receive feedback.

Instructions:

1. Read Steps 1, 2, and 3.
2. Answer Steps 4 and 5 in Spanish or English.
3. Answer Step 6 in Spanish.
4. Use the image to make the necessary changes so the joke is funny in English.

Step 1: Source Text:



UNA OPTIMISTA

—¡Me río de lo tontísimo que es mi marido! Figúrese que acabo de verlo en un taxi con una.
—¡No sería él!
—Sí, hombre, sí, ¡Es que el se figura que yo no sé que estamos a 28 de diciembre!

327

Una optimista
- ¡Me río de lo tontísimo que es mi marido! Figúrese que acabo de verlo en un taxi con una.
- ¡No sería él!
-Sí, hombre, sí. ¡Es que él se figura que yo no sé que estamos a 28 de diciembre!

Step 2.1: GT Output
An optimist
- I laugh at how stupid my husband is! Imagine I just saw you in a taxi with one.
- It wouldn't be him!
- Yeah, man, yeah. It's just that he figures that I don't know that it's December 28!

Step 2.2 GT Back-Translation
Un optimista
- ¡Me río de lo estúpido que es mi esposo! Imagina que te acabo de ver en un taxi con uno.
- ¡No sería él!
- Sí, hombre, sí. ¡Es que se imagina que no sé que es 28 de diciembre!

Step 3.1 DT Output
An optimist
-I laugh at how silly my husband is! Imagine I just saw him in a cab with one!
-It wouldn't be him!
-Yes, man, yes, it's just that he thinks I don't know it's December 28th!

Step 3.2 DT Back-Translation

Un optimista
- ¡Me río de lo tonto que es mi marido! ¡Imagínate que lo acabo de ver en un taxi con una!
- ¡No sería él!
- ¡Sí, hombre, sí, es que cree que no sé que es 28 de diciembre!

Step 4: AI Model Creative Translation

Chat with Gemini, ChatGPT, and Copilot to find a creative translation for the original text. Use strong and varied prompts to obtain better results.

Copy and paste your chat with Google Gemini Pro (https://gemini.google.com/app)

Copy and paste your chat with ChatGPT (https://chatgpt.com/)

Copy and paste your chat with Copilot (https://copilot.microsoft.com/)

Step 5: Comparison and Creative Translation

Compare all tools (MT services and AI models) used and identify a creative English translation.

- Do these comparisons help you understand the Spanish text? Explain.
- Write your creative English translation. Remember, a literal translation is not acceptable.

Step 6: Reflection (in Spanish)

Write a paragraph in Spanish reflecting on the challenges encountered when translating or improving the text given by GT, DT, Gemini, ChatGPT, and Copilot. Explain the problems you found and how you solved them. Describe the translation method used (free or interpretative-communicative) and justify your choice.

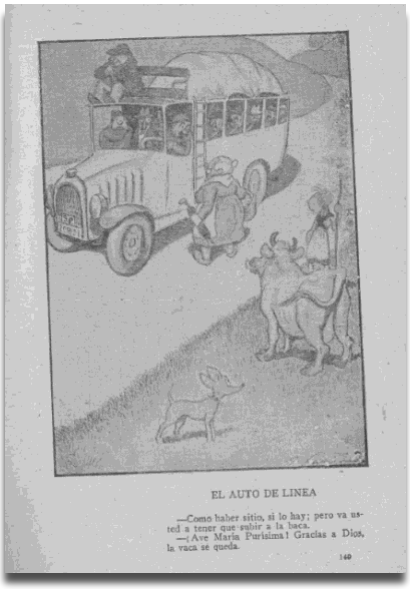
Rubric for Translation Activity 1 (15 Points)

Criteria	Points	Description
Step 4: Creative Translation Prompt and Results Obtained	4	Develops strong and effective prompts for Gemini, ChatGPT, and Copilot, taking into account the context and nuances of the joke. The results effectively enhance both the humor and contextual relevance of the translation prompt.
Step 5: Creative Translation	3	Provides a creative and humorous English translation that captures the original joke's intent and effect.
Step 5: Comparison of Tools	4	Effectively compares outputs from all tools, discussing strengths/weaknesses and handling of humor/context.
Step 6: Reflection	4	Thoughtful reflection in Spanish on the translation process, challenges, solutions and method solutions. Explains how the image influenced the translation decisions.

Translation Activity 2

Note: Translation activity 2 uses the same instructions and rubric as Activity 1. Only Steps 1-3 differ.

Step 1: Source Text:



El auto de línea
- Como haber sitio sí lo hay; pero va usted a tener que subir a la baca.
- ¡Ave María Purísima! Gracias a Dios, la vaca se queda.

Step 2.1: GT Output
The bus line
- As if there is room; but you're going to have to get on the roof.
- The Virgin Mary! Thank God, the cow stays.

Step 2.2: GT Back-Translation
La línea de autobús
- Como si hubiera lugar; pero vas a tener que subirte al techo.
- ¡La Virgen María! Gracias a Dios, la vaca se queda.

Step 3.1: DT Output
The line car
- There is room, but you'll have to climb on the roof rack.
- Ave Maria Purisima! Thank God, the cow is staying.

Step 3.2 DT Back-Translation
El coche de línea
- Hay espacio, pero tendrás que subir a la baca.
- ¡Ave María Purísima! Gracias a Dios, la vaca se queda.

Appendix B

Step 4: Translation Task Outputs

Student 1

ChatGPT

Prompt:
Translate this joke to make sense in English
ChatGPT output:
Sure, I'd be happy to help! Just provide me with the joke you'd like translated, and I'll do my best to make it work in English.
Prompt:
- Como haber sitio sí lo hay; pero va usted a tener que subir a la baca.
- ¡Ave María Purísima! Gracias a Dios, la vaca se queda.
ChatGPT output:
Here's a translation of the joke that should make sense in English:

- Is there room?
- There is, but you'll have to climb onto the roof.
- Oh, thank goodness! Thank God the cow is staying!

In this joke, the humor comes from the confusion between "baca" (which means "roof rack" in some Spanish-speaking regions) and "vaca" (which means "cow"). The person misunderstands "baca" for "vaca" and is relieved to hear that a cow isn't involved.

Prompt:
What is an English equivalent of "Ave María Purísima" that is not a direct translation and that an American audience can understand?
ChatGPT output:
An English equivalent of "¡Ave María Purísima!" that would resonate with an American audience might be:
"Good heavens!" or "Oh my goodness!"
These expressions capture the surprise or relief typically conveyed by "¡Ave María Purísima!" in a way that fits naturally into English conversation.

Prompt:
Translate this joke into English for an American audience
Gemini output:
Please provide the joke you'd like me to translate.
Once I have the joke, I'll do my best to adapt it for an American audience while preserving its humor.
Here are some things I'll consider when translating:
● Cultural references: I'll ensure that any references are understandable to Americans.
● Wordplay and puns: I'll try to find equivalent wordplay or puns in English.
● Humor style: I'll adapt the humor to match American sensibilities.
Please feel free to share the joke, and I'll get started.
Prompt:
- Como haber sitio sí lo hay; pero va usted a tener que subir a la baca.

- ¡Ave María Purísima! Gracias a Dios, la vaca se queda.

Gemini output:

¡Claro que puedo ayudarte a traducir ese chiste! Aquí tienes un par de opciones para adaptarlo a un público estadounidense, manteniendo la esencia del humor:

Option 1: More direct wordplay

Original: Como haber sitio sí lo hay; pero va usted a tener que subir a la baca. ¡Ave María Purísima! Gracias a Dios, la vaca se queda.

Translation: "Well, there is room, but you'll have to ride on the roof rack. Holy cow! Thank goodness the cow stays."

Explanation: Here, the joke plays on te word "baca" (roof rack) and the confusion with "vaca" (cow). The relief that the cow stays is an exaggerated humorous touch.

Gemini

Option 2: More visual and absurd humor

Original: Como haber sitio sí lo hay; pero va usted a tener a que subir a la baca. ¡Ave María Purísima! Gracias a Dios, la vaca se queda.

Translation: "Sure, there's room, but you'll have to sit on the cow. Holy cow! Thank goodness we can leave the roof rack behind."

Explanation: This option exaggerates the situation, imagining the passenger riding the cow instead of the typical seat on the car roof.

Additional considerations:

● Context: For a more precise translation, it would be helpful to know the context in which the joke is told. Is it a conversation between friends? Is it happening on a road trip?

● Audience: The level of absurd humor can be adjusted according to the intended audience.

● Cultural equivalent: if there is an equivalent expression or situation in American culture that can substitute "Ave María Purísima," it would be ideal to incorporate it.

Other options:

● "Well, there is room, but you'll have to sit on the cow. That's one way to moo-ve!" (Pun with "move" and the sound of a cow)

● "Sure, there's room, but you'll have to ride on the roof. Thank goodness we don't have to milk this situation any further." (Cow-related pun)

Prompt:

translate this joke into English for an American audience

-Como haber sitio sí lo hay; pero va usted a tener que subir a la baca.

- ¡Ave María Purísima! Gracias a Dios, la vaca se queda.

Copilot

Copilot output:

Sure! Here's a translation of the joke for an American audience:

● "Well, there is space, but you'll have to ride on the roof rack."

● "Holy Mary! Thank God, the cow stays."

It's a play on words where "baca" (roof rack) is confused with "vaca" (cow). Hope that brings a smile to your face! 😁

<u>Student 2</u>

Prompt:

ChatGPT	Traduce este texto al inglés comprendiendo el chiste que contiene "Como haber sitio sí lo hay; pero va usted a tener que subir a la baca. - ¡Ave María Purísima! Gracias a Dios, la vaca se queda." **ChatGPT output**: Aquí está la traducción al inglés manteniendo el chiste: "As for space, there is some; but you're going to have to ride on the roof rack (baca). - 'Holy Mary! Thank God, the cow stays.'" *El juego de palabras ocurre porque "baca" (que es la parrilla del techo de un vehículo) se confunde con "vaca" (el animal), lo que genera la respuesta cómica.*
Gemini	**Prompt**: Traduce este texto al idioma de inglés, comprendiendo lo que conlleva que sería el chiste del mismo "Como haber sitio sí lo hay; pero va usted a tener que subir a la baca. - ¡Ave María Purísima! Gracias a Dios, la vaca se queda." **Gemini output**: ¡Claro que sí! Vamos a traducir el chiste y a explicar un poco el humor detrás de él: **Literal translation**: "As for there being room, yes there is; but you'll have to climb onto the cow. - Holy Mary, most pure! Thank God, the cow stays." **More natural translation with the joke explained**: "If you're asking if there's room, yes, there is; but you'll have to ride on the roof rack. - Holy Mary, most pure! Thank God, the roof rack stays." **Explanation**: The humor in this joke lies in the wordplay and exaggeration. The person asking if there is space in a vehicle receives an absurd answer: that they must get on the cow (or on the car roof, in the more common translation). The second part of the joke, "¡Ave María Purísima! Gracias a Dios, la vaca se queda.", is an exaggerated expression of relief, as if the person was about to do something dangerous and is relieved not to have to do it. **English translation capturing the humor**: "Are there any seats left? Sure, but you'll have to ride on the roof! - Thank goodness! I'm not getting up there!" **Another, more literal option:** "As for there being space, yes there is; but you'll have to climb on the cow. - Holy Mary, most pure! Thank God, the cow stays put."
Copilot	**Prompt**: Traduce este texto al inglés, teniendo en mente el chiste que contiene "- Como haber sitio sí lo hay; pero va usted a tener que subir a la baca. - ¡Ave María Purísima! Gracias a Dios, la vaca se queda." **Copilot Output**: ¡Claro! Aquí tienes la traducción, manteniendo el chiste: "Well, there is space; but you'll have to climb up to the roof rack. -Holy Mary! Thank God, the cow stays." Espero que el chiste se mantenga igual de divertido en inglés. ¿Hay algo más en lo que pueda ayudarte? 😊

About the Author

Cristina Pardo-Ballester (Ph.D., University of California, Davis 2007; M.A. University of Nevada, Reno 1999) is an Associate Professor of Hispanic Linguistics in the Department of World Languages and Cultures at Iowa State University. She is a linguist with broad interdisciplinary interests in Spanish applied linguistics and beyond. Her research interests center on translation studies, Spanish culture, intercultural communication, second language acquisition, technology-enhanced language learning, and language assessment. She has published in Language Learning & Technology, Hispania, Foreign Language Annals, CALICO Journal, Language Assessment Quarterly, etc. She has also published her work in books, newsletters, and the proceedings of international conferences. She has co-authored a book, *Pura Vida* (2014 & 2020), to promote cross-cultural knowledge. She has recently published works on the impact of technology on Spanish learners to promote intercultural competence using Spanish jokes in which historical context is key for understanding humor.
www.cristinapardoballester.com

Chapter 8

Indonesian Learners' Perceptions on Human-AI Interaction in Speaking Assessments

Andrias Susanto
Iowa State University

This study investigates Indonesian EFL learners' perceptions of generative AI (GenAI) applications compared to their perceptions of human tutors in an online test preparation course. Semi-structured interviews with 11 participants who had participated in practice with both a GenAI tool and a human revealed that the GenAI tools were valued for their efficiency, objectivity, and detailed feedback, particularly in pause detection and word count analysis. However, learners highlighted the importance of human tutors for personalized, adaptive, and in-depth feedback. They saw GenAI as providing consistent and rapid scoring, but human tutors as fostering a more supportive learning environment through individualized guidance. Many participants favored a hybrid approach, integrating GenAI's analytical precision with human interaction to enhance test preparation and real-world speaking skills. Notably, AI was generally trusted more for assessment consistency, whereas human tutors were preferred for instructional support. These findings suggest that for EFL learners in a test preparation course a combination of GenAI and human tutors may be the most satisfying.

Cite as: Susanto, A. (2025). Indonesian learners' perceptions on human-AI interaction in speaking assessments. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 145-165). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.08
©2025 Susanto. Published under a CC BY license.

Introduction

One significant challenge of learning English in countries where it is a foreign language (EFL), like Indonesia, is the limited opportunities for learners to regularly practice their speaking skills with another individual, get a score estimate, and feedback (Celce-Murcia et al., 2014). To do this, EFL learners often need to enroll in courses specifically designed to develop their English speaking skills (Thornbury, 2005). Moreover, with the expansion of the Internet, learners have increasing opportunities to access online English courses, connecting teachers and students across various time zones and locations (Zhu, 2023). For instance, a teacher based in the North American Midwest region can conduct lessons for learners throughout areas of Indonesia, e.g. Sumatra, Java, Bali, Borneo, Sulawesi, and Papua either synchronously or asynchronously.

English tutors play a vital role beyond instruction by inspiring learners and providing detailed speaking assessment of learners' progress as they develop their speaking. Human tutors evaluate pronunciation, grammar, and fluency, offering personalized, culturally appropriate feedback. By actively engaging with students to clarify misunderstandings and suggest improvements, tutors create a supportive environment that boosts confidence and fosters language development (Susanto, 2018). Nonetheless, one common challenge in human-to-human tutoring is coordinating schedules between the tutor and the learner, as tutors are not always available around the clock. This can make it difficult to find mutually convenient times. Additionally, class meeting hours are often fixed, which limits opportunities for in-class practice and hampers the promotion of independent speaking activities outside of scheduled sessions. As human tutors are not always available when learners need immediate feedback, learners often submit speaking recordings to their tutors via online platforms, such as email or learning management systems (e.g., Google Classroom, Canvas, Moodle, Blackboard, etc.), allowing tutors to review them asynchronously when their schedules permit. While this approach offers flexibility, it may not be ideal for learners preparing for English-speaking tests, where real-time interactive practice with immediate feedback is crucial.

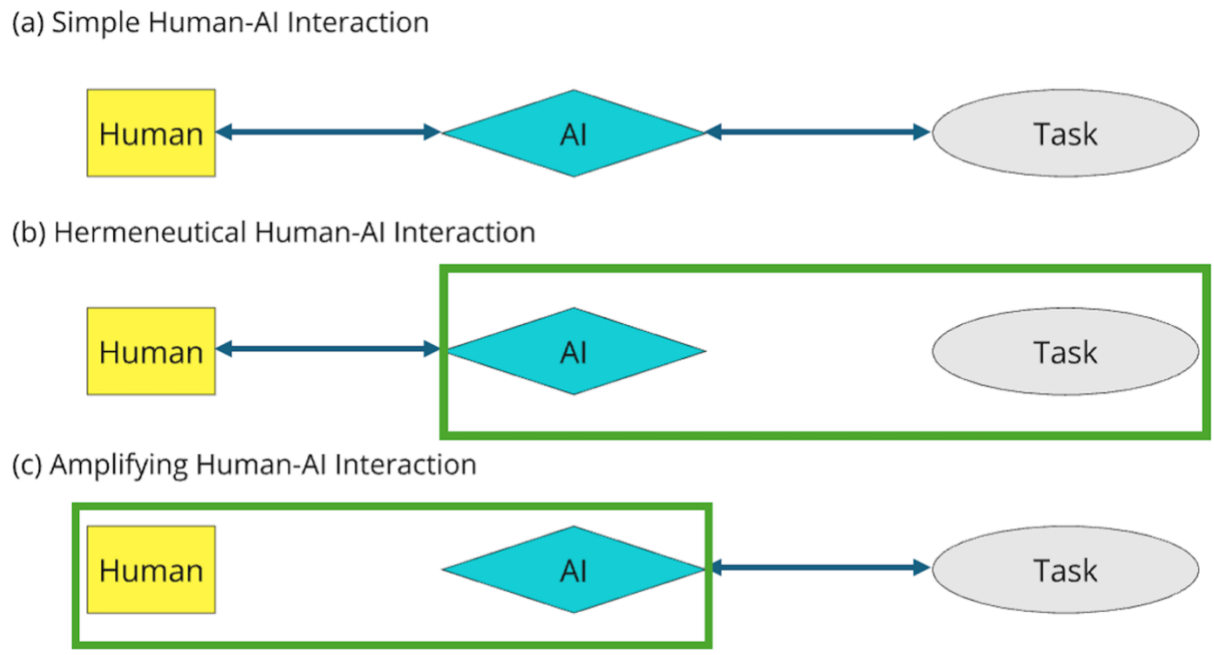
To address this challenge, human-computer interaction (HCI) can utilize AI-powered applications to enhance language learning by allowing learners to engage in simulated conversations with a generative AI (GenAI) that assesses speech and provides real-time feedback through its language processing capabilities. Such tools are created with technologies that generate a hyper-realistic three-dimensional character that resembles a human tutor who "speaks" using synthetic audio (Pataranutaporn et al., 2021) and processes learners' language using automated speech recognition to evaluate pronunciation, fluency, and grammar, offering instant scoring and personalized corrections (Isaacs, 2018). By adapting to learners' responses and tracking progress, such GenAI tools enable independent practice as well as interactive and structured learning experiences. Certain roles traditionally filled by teachers—such as prompting questions, providing score estimates, and offering feedback—can now be performed by machines. Moreover, with features like audio playback for performance review, these tools may help learners identify and improve specific weaknesses, particularly in fluency and pronunciation. By offering a flexible and accessible platform, this human-AI interaction (HAI)

technology may appear to meet the need for regular speaking practice and score estimation, available anytime and anywhere learners choose to engage with it.

Hollnagel (in Rasmussen & Andersen, 1991) identified three types of HAI: simple, hermeneutical, and amplifying (Figure 8.1). Simple HAI refers to task-based interactions where AI functions as a basic tool to enhance efficiency. Hermeneutical HAI, on the other hand, focuses on aiding users in interpreting complex data, with AI acting as a guide. Amplifying HAI involves collaboration between humans and AI to enhance outcomes. These different types of HAI suggest different ways that AI may empower learners and teachers to achieve more than traditional human-human tutoring for speaking test preparation (Sharples & Pérez, 2022). The three AI uses describe the ways that AI can serve as a tool–much like a ladder enables individuals to reach greater heights. For instance, in this context, SpeechAce, the website that provides these AI services used in this study, serves as an amplifying tool by allowing independent test administration, scoring, and feedback, particularly when participants experience constraints of time or location. Amplifying HAI also suggests a collaboration between human tutors and AI to improve the ratings by reporting of total number of words and length of pauses per minute in addition to the judgments made by the human raters.

Figure 8.1

Three Types of Human-AI Interaction (adapted from Hollnagel in Rasmussen & Andersen, 1991)



In other words, given the distinctions among the three types of relationships between human, AI, and task, this present study is aimed at exploring amplifying HAI in speaking assessment by investigating how humans and AI technology work together to complete an assessment task. In the context of technology-mediated language assessment (Chapelle & Voss, 2016; Sadeghi & Douglas, 2023), the use of technology in the design of the test method

necessitates research. When technology is used to create a test method with amplifying AI, it is crucial to distinguish between the constructs measured by the human-delivered and scored speaking test versus the computer-delivered and scored speaking test, as the two test methods may not necessarily assess the same constructs. Even though a detailed exploration of the differences between human-delivered and computer-delivered speaking tests lies beyond the scope of this study, it is important to note that they may assess different constructs.

In this study the primary concern is the use of amplifying AI to assist EFL learners test preparation for taking their International English Language Testing System (IELTS) (Read, 2022) speaking test. In this context, the assessment must be evaluated for the utility of the feedback that it provides to learners. Recent research, such as Woo & Choi's (2021) systematic review, highlights the potential of AI-powered language learning tools that support natural language processing to provide personalized, adaptive learning experiences with instant feedback. These tools are proposed to enhance language proficiency by tailoring lessons to individual needs, offering real-time corrections, and adjusting difficulty based on learner progress. While promising, challenges remain regarding fairness (Chapelle, 2013; McNamara, 2019)—such as the risk of bias in AI systems—and the need for further research into the effectiveness of these technologies across diverse learner groups. Particularly in relation to AI as a conversational partner, researchers (Loukina & Buzick, 2017; Macinska & Vinkler, 2024; Van Moere & Downey, 2016) argue that AI-based tools can also play a key role in developing interactional competence, which is vital for L2 acquisition. Interaction provides opportunities for output, facilitates input and feedback, and encourages learners to negotiate or clarify meaning. One conclusion that can be drawn from the above research is that AI-powered language learning tools may have significant potential to improve language proficiency by offering personalized, adaptive learning experiences.

Fully automated scoring and feedback systems for English speaking assessments claim benefits like efficiency, fairness, and time savings for teachers (Isaacs, 2018; Khabbazbashi et al., 2021; Pearson, 2023). These systems can handle many assessments at once and reduce grading time, allowing educators to focus on other tasks. Moreover, studies show that automated scores often match human scores closely (Xi et al., 2024), and immediate feedback helps students identify areas for improvement (Zechner et al., 2015). However, these systems may discourage creative responses and sometimes give higher scores to lower-proficiency speakers (Zechner & Evanini, 2019). While automated scoring can motivate independent practice, it may also lead students to focus on test performance rather than real-life communication. The feedback, although fast, may lack the personalization and depth that human feedback offers (Isotalus et al., 2024). Therefore, to make the most of these systems, teachers should help students use them to improve their overall language skills, not just for test preparation.

This study explores AI-powered test preparation for a speaking test by gathering insights from EFL learners in Indonesia. It assessed the learners' perceptions of the quality of the AI-generated video prompts, automated scoring, and feedback. Since learners' perceptions play a key role in their success, especially in online learning (Martin et al., 2020; Schunk & Meece, 1992), this research asks the learners directly about their perceptions of the AI relative to their

experience with human tutors for supporting their English-speaking practice. The research question is "How do learners perceive AI-generated video prompts, automated scoring, and feedback provided by AI compared to the prompts, scoring, and feedback they receive from the human tutors?"

Methods

In this qualitative study, Indonesian EFL learners who had completed a month-long IELTS preparation course with a human tutor were invited to share their perspectives on the use of the AI-powered platform SpeechAce (https://www.SpeechAce.com/) in comparison to their experience with the human tutor. To elicit learners' comparative judgments, semi-structured interviews were employed, and the learners' responses were analyzed to identify themes pertaining to their comparative judgments of SpeechAce and the human tutors as interlocutors for their IELTS Speaking test preparation.

Context and Participants

The participants, aged between 18 and 30, were alumni of an online IELTS preparation course that included weekly speaking practice sessions. Invitations to participate were sent to the entire group immediately after they had completed the post-test for the course. Although 45 alumni were invited, only 11 volunteered, a response rate of 24%. Five participants were interviewed in August 2024, while the remaining six were interviewed in September 2024. The former group had taken the course in July, and the latter in August 2024. Therefore, it can be assumed that their speaking proficiency was approximately the same as when they were actively engaging with their tutors. These participants, whose English proficiency levels ranged from B1 to B2 on the CEFR scale, had taken part in speaking test simulations with their tutors as part of the post-test administered at the end of the course.

The tutors were native Indonesian speakers who had previously taken the IELTS test and achieved an overall band score of 7.0. Some of them already held a master's degree from overseas universities in various fields, including Educational Technology and Teaching English to Speakers of Other Languages (TESOL). The course is a privately-run IELTS preparation program that follows a specific curriculum designed by the institution itself. All tutors had completed an official online IELTS teacher training program and had also undergone internal training.

Human Tutoring and Feedback

Before volunteering for the study, each student had participated in at least two one-on-one speaking test simulations with the tutor, each lasting about 15 minutes. During these sessions, the tutor took on the role of the IELTS speaking examiner, asking questions that the learners answered according to the guidelines they had learned in class. Following each simulation, the tutor provided feedback based on the four criteria outlined in the IELTS speaking rubric (https://takeielts.britishcouncil.org/sites/default/files/ielts_speaking_band_descriptors.pdf) and offered specific comments, as illustrated in Table 8.1.

Table 8.1

Example Tutor's Notes Showing Typical Error Feedback that a Student Might Receive

IELTS Practice Test Part	Tutor's Notes from Angii's Record
Part 1	'uh...' fillers are present throughout the answers when you are become --> when you become compare when I was --> In comparison to my previous ... I still thinking about --> I am still thinking about some pauses are still present /own/ mispronounced every weekends --> weekend
Part 2	'uh...' fillers are present throughout the answers go around to the neighborhood --> go around the neighborhood the computers still only less program --> the computer had fewer computer programs /game/ mispronounced I addict to that game --> I got addicted to that game I told that the computer --> I said that the computer the technology become --> has become
Part 3	'uh...' fillers are present throughout the answers As my experience --> From my experience, self-corrections and repetitions are still present (may lead to score reduction) I support with the school --> I support the school For the teachers taught the children --> for the teachers to teach the students independent to the teacher --> independent to learn not depending on the teacher The information in the computer was made by humans --> is made /made/ mispronounced the knowledge was taught --> is taught on internet --> on the internet without the teachers itself --> with no teacher at all.
Total IELTS Speaking score estimate	5.0

In Part 1 the tutors might comment that several issues had affected this learner's speaking fluency and accuracy. The use of fillers like "uh..." disrupted the flow of speech and lowered the fluency score. There were grammatical errors, such as "when you are become" (should be "when you become") and "I still thinking about" (should be "I am still thinking about"). The phrase "compare when I was" could be rephrased as "In comparison to my previous..." to sound more natural. Pauses in speech were also noted, affecting fluency. Pronunciation issues included "own," which was mispronounced, and "every weekends," which should be "weekend."

In Part 2, typical comments would identify similar issues. Fillers like “uh...” again disrupted fluency. Grammatical mistakes included “go around to the neighborhood” (should be “go around the neighborhood”) and “the computers still only less program” (should be “the computer had fewer computer programs”). The word “game” was mispronounced, and “I addict to that game” should be “I got addicted to that game.” The phrase “I told that the computer” should be “I told my father that,” and “the technology become” should be “has become.”

In Part 3, the tutors might find issues with fluency, coherence, and grammar. The use of fillers like “uh...” disrupted the flow of speech. Phrasing errors included “As my experience” (should be “From my experience”) and “I support with the school” (should be “I support the school”). Self-corrections and repetitions may also lower the fluency score. Some sentences were awkward, such as “independent to the teacher” (should be “independent to learn, and not depending on the teacher”). Tense errors included “The information in the computer was made by humans” (should be “is made”) and “the knowledge was taught” (should be “is taught”). Pronunciation issues included the word “made,” and “on internet” should be “on the internet.” The phrase “without the teachers itself” should be “with no teacher at all.”

After explaining each error and suggesting the correct forms, the tutor provided a score estimate for the learners as shown in the example using the scoring rubric of the IELTS Speaking section that is available on the official website. The possible score range is based on the official IELTS band scale, which goes from Band 1 (non-user) to Band 9 (expert user). Following this, the tutor shared techniques to help avoid similar mistakes in the future. For example, the tutor advised that while pauses are acceptable, they should be placed at the end of a sentence. Learners were also encouraged to practice reading aloud to improve their understanding of thought groups, helping them know where to place pauses and stops appropriately. This feedback session lasted about 5 minutes, and unless the learners requested further clarification, the tutor moved on to the next student. However, before moving on, the learners were asked to confirm their understanding of what they needed to work on and the practical steps they should take to improve their overall speaking performance.

SpeechAce: AI-Powered Speaking Evaluator

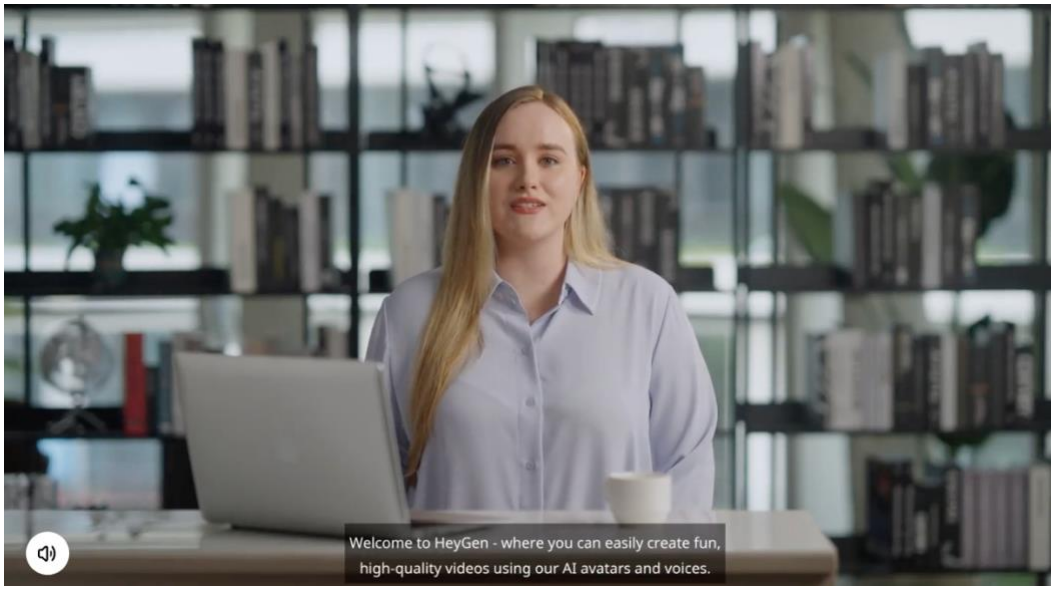
SpeechAce is an online platform that integrates advanced generative AI technologies to create a hyper-realistic humanoid avatar and video-audio generator as well as natural language processing tools such as speech recognition specifically designed to assess pronunciation and fluency. This study focuses on three key features of SpeechAce: (1) AI-generated video content featuring a speaking interlocutor, (2) score estimates provided by the system, and (3) the feedback given to learners.

The video displays an avatar examiner, generated by AI through a synthesis process. This process combines various digital AI generated elements from a stock of digital personalities to create lifelike avatars that closely mimic human appearance and behavior (Schiefelbein, 2023). These avatars can be used in various applications, from customer service to language assessment for personalized interactions or as virtual influencers to engage with followers in a natural and

relatable manner. Platforms like Synthesia (https://www.synthesia.io/) and Heygen (https://www.heygen.com/) are commonly used by individuals and enterprises to create these humanoid video avatars. These avatars can read aloud scripted texts input by learners and also utilize different audio options, offering a wide range of human voices and accents based on learner preferences (Bougueffa et al., 2024).

Figure 8.2

A Snapshot of a Humanoid Hyper-Realistic Avatar Video (from https://app.heygen.com/signup)



The scoring system was developed from the speech recognition system by comparing learner performance with established standards. These standards help create a classification system for evaluating speaking performance. According to the website, the scores generated are interpretable on the Common European Framework of References (CEFR) and are comparable to scores provided on standardized tests like IELTS, PTE, TOEFL, and TOEIC speaking sections. The scores are said to be scientifically tested, since they have been evaluated using thousands of speaking samples from real test takers and compared against blind evaluations by professional test examiners. The developers also claim that the system delivers scores within 0.5 to 1.0 points of those awarded by human examiners and has an assessed test-retest reliability coefficient of 0.82 (Pearson Correlation Coefficient). Table 8.2 captures the snapshots of the score comparisons against the other tests for a score of 7.1 on SpeechAce.

Table 8.2

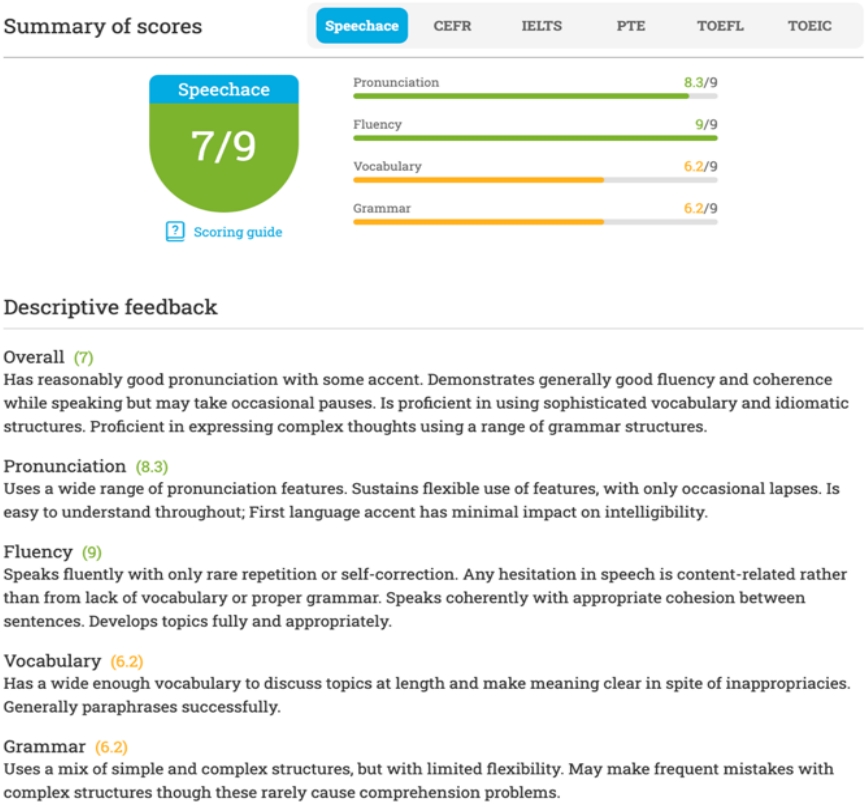
An Example of SpeechAce Score 7.1 out of 9.0 Interpreted on the CEFR and in Comparison to Standardized Tests (from https://www.SpeechAce.com/speaking-test/)

Tests	Scores
IELTS	7.0 out of 9.0
PTE	66 out of 90
TOEFL	23 out of 30
TOEIC	920 out of 990

The feedback feature of SpeechAce is divided into two parts: descriptive feedback per criterion and per question. For the former, the website provides a general score description of the overall performance and specific feedback for each criterion, including pronunciation, fluency, vocabulary, and grammar, as shown in Figure 8.3. For example, the Pronunciation criterion is described as: "*Uses a wide range of pronunciation features. Sustains flexible use of features, with only occasional lapses. Is easy to understand throughout; First language accent has minimal impact on intelligibility*." This description corresponds to a score of 190 for TOEIC, 29 for TOEFL, 87 for PTE, and 8.5 for IELTS, leading to a categorization of C2 in the CEFR.

Figure 8.3

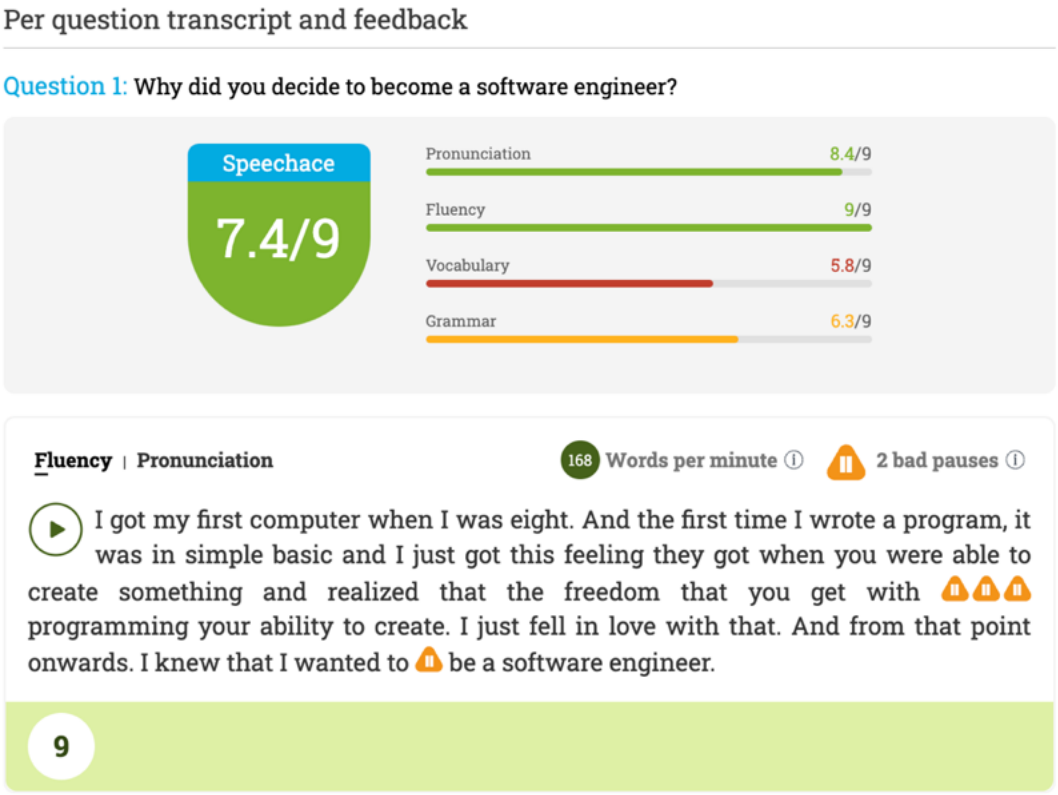
Descriptive Feedback Given Based on the Learners' Speaking Performance



For the question specific feedback, the website displays each of the three questions, the learners' responses for each question, and analyses of the total number of words per minute, and location of bad pauses (represented by an orange triangle with two vertical stripes in it in Figure 8.4).

Figure 8.4

Example Question-Level Feedback Provided to A Learner



The Semi-Structured Interview

The interview questions were designed to elicit participants' perceptions of three key aspects of the AI technology: (1) the learners' general perceptions of AI-generated humanoid video prompts compared to direct test simulations delivered by the human instructor, (2) the perceived accuracy of automated scoring versus human scoring, and (3) their perceptions on the clarity of the feedback provided by the system in comparison to feedback from human instructors. The interviews were conducted in a conversational style to encourage participants to share their genuine thoughts and experiences, with the goal of capturing nuanced perspectives, including any hesitations or concerns about relying on AI in speaking assessments. The participants could choose to speak in either Indonesian or English, depending on their preference and convenience.

Data Collection Procedures

Upon receiving Institutional Review Board (IRB) approval (IRB 24-380), I distributed an online survey to prospective participants to describe the study and obtain their consent to participate. Participants' written consent and email addresses were collected to arrange interview schedules and to conduct user experience sessions and follow-up interviews.

The participants who agreed to participate were briefed on the research objectives and the role of AI technologies in speaking assessment. During the session, SpeechAce was introduced, and participants received brief step-by-step guidance on how to navigate the site and use its features. This included instructions on accessing AI-generated video prompts, submitting spoken responses, and reviewing the automated feedback and scoring provided by the system. Each participant was then given approximately 10 minutes to independently work on the platform, interacting with AI-generated prompts that mimicked typical IELTS Speaking questions, and completing a set of speaking tasks that consisted of three questions with one-minute-long responses. After recording their responses, participants received automated scores and feedback, which included detailed insights into their pronunciation, fluency, and areas for improvement. Following their interaction with the platform, they participated in interviews with me via Google Meet, which had also served as a medium to introduce participants to SpeechAce.

The time allocated for each interview was between 20 and 30 minutes. Each of the 11 interviews was conducted on a different day in August and September 2024, with a total of 11 days for completion. The interviews were conducted one-on-one via video conference and were mainly in English, although occasionally involved the use of Indonesian, which the researcher, a native Indonesian speaker, helped translate into English. Audio recordings were made of the interviews to ensure accuracy and were later transcribed. If any responses were in Indonesian, they were translated by the researcher into English.

The semi-structured interviews were audio-recorded and transcribed verbatim to ensure the accuracy of the data and preserve the original context of participants' responses. All interview transcripts were collected in separate files and subsequently organized for thematic analysis. The transcribed interviews yielded 46,332 words, excluding the interview questions, across a total of slightly less than 6.5 hours of interview time. The shortest interview lasted 21 minutes, while the longest lasted 28 minutes. On average, each participant produced 4,212 words in response to the three main questions, with individual word counts ranging from 2,552 to 4,254 words.

Data Analysis

Once the data were transcribed, an initial review was conducted manually to identify the most frequently mentioned themes and to understand the context in which these themes emerged. For example, participants expressed varying preferences regarding AI-generated video prompts compared to human instructor feedback. The category of "preference for human feedback" was

identified, and then the frequency of occurrence of this theme across the eleven responses was noted.

Guided by the interview prompts, three overarching themes were identified: (1) test administration, (2) scoring, and (3) feedback. These themes served as analytical categories to compare participants' experiences with the AI-powered SpeechAce platform and the human tutor. The analysis focused on how learners articulated their perceptions, preferences, and concerns in relation to each theme. In addition, patterns were identified regarding perceptions about the effectiveness and clarity of AI versus human feedback and assessing learners perceived usefulness of the platform for improving their speaking performance. To ensure accuracy and maintain participants' voices, excerpts were extracted verbatim from the transcripts and categorized accordingly. To ensure participant confidentiality, codes (P1 to P11) were used as unique identifiers for the respondents.

Results

These themes highlighted the differences rather than similarities between the two modalities—the human tutor and SpeechAce. Participants expressed contrasting experiences in how each approach handled test administration, feedback delivery, and scoring. Despite the focus on these differences, most learners emphasized the value of combining both approaches, suggesting that human and AI feedback could complement each other to enhance the overall learning experience.

Learners' Perceptions of Test Simulation Administrations

Learners' perceptions of test simulation administrations using AI-generated video prompts compared to those delivered by human instructors reveal a unique view of the strengths and limitations of both approaches. Many learners express a preference for human tutors due to the personalized, real-time feedback they provide, particularly in assessing non-verbal communication such as tone, facial expressions, and body language, which are critical for effective communication. Participant 3 explained:

> I think I prefer the human one because the AI still feels a bit awkward, and it makes me more nervous because I can't read the examiner's body language. When doing the test in SpeechAce, I felt kind of overwhelmed because of the timer sound. With a human tutor, the timer doesn't really have a sound and they only warn us when the time is up, so it helps me more to focus on thinking about the content I want to give as an answer rather than focusing on the limit. (P3)

Practicing speaking with a human tutor is often regarded as more natural, as language fundamentally serves as a medium of communication between people. While SpeechAce can mimic human interaction through its interface and speech capabilities, it remains an artificial as an interlocutor for spoken language. Some learners view it as a distinct and less authentic

experience, which could affect their speaking performance. From this standpoint, the interaction facilitated by SpeechAce can be considered less life-like, as expressed by Participant 7.

> I prefer speaking practice with an IELTS tutor because it feels more real-time and lively. Using SpeechAce might be helpful, but only as a tool for self-practice. (P7)

On the other hand, Participant 2, who saw an advantage to the AI tool, found it more comfortable practicing without the pressure of human interaction:

> I feel less nervous when practicing speaking with SpeechAce because I don't have to interact with someone directly. Therefore, as a beginner, I feel more comfortable using SpeechAce. (P2)

Other learners advocate for combining both AI and human support systems, recognizing that AI tools can complement the real-time, dynamic interaction of human tutors, creating a more comprehensive and effective learning experience. Participant 10 was among those who recognized the complementarity of the two:

> In my opinion, using SpeechAce for speaking practice helps me build confidence in speaking. The app provides structured exercises and useful feedback. However, to create a practice atmosphere that closely resembles real-life situations, speaking directly with a tutor remains a more effective approach. Direct interaction with a tutor allows me to learn how to respond spontaneously and receive more personalized corrections. This way, I can develop my speaking skills more effectively and overcome communication challenges I might face in everyday situations. Both SpeechAce and tutors complement each other." (P10)

The responses suggest that participants see practicing with a machine is inherently different from working with a human tutor. Learners can sense this distinction, which may influence their performance. Eight out of the 11 participants (72.73%) argued that experiencing both methods was beneficial, as they complement each other. Practicing with a computer can help with practicing with a limited level of nervousness, while speaking with a real human tutor is essential for preparing for the actual speaking test. The other three participants expressed differing preferences: two stated that they preferred the human-administered exam, while one found the computer-administered version to be much more preferable.

Learners' Perceptions on Scoring

Participants' perceptions revealed a distinct divide between those who favor the objectivity and efficiency of AI-generated feedback and those who value the personalized, nuanced assessments offered by human tutors. Five participants highlighted their preference for computer feedback in the form of the quick and standardized scoring provided by SpeechAce. They expressed appreciation for the app's ability to deliver immediate scoring, enabling them to efficiently track their progress and compare performance across different tests. For instance, P9 emphasized how SpeechAce's rapid and objective evaluations, including detailed metrics like words per minute to

determine fluency level and assist in monitoring development in specific areas, like fluency. While human tutors might offer similar assessments, doing so in real-time is not feasible, as it would require them to simultaneously and repetitively calculate words per minute and monitor speech pauses tasks that are better handled by automated systems. Participant 9 expressed appreciation for the rapid scoring provided by SpeechAce:

> To get a score, I prefer using SpeechAce because the results are obtained more quickly and are standardized. This app provides objective assessments and directly shows the evaluation results clearly. Additionally, SpeechAce also displays comparisons across various types of tests, so I can see my progress from different aspects and compare my performance in different exercises. With features like this, I find it easier to track the development of my speaking skills numerically and identify areas that need specific improvement. The speed and accuracy of the results make me feel more comfortable in preparing for the speaking exam. (P9)

However, there was also significant support for human tutors, with nine participants (81.82%) highlighting the perceived accuracy of the score they received. Only two participants stated that they preferred the feedback from SpeechAce, considering it more objective since it is programmed to generate specific scores based on the computer's analysis. The nine participants emphasized that they prefer the scores given by human tutors because they believe these scores reflect a more real assessment of their speaking ability. P8 also mentioned that human interaction allows for clarification, which is not always possible with AI. Additionally, P11 indicated that while SpeechAce provides thorough feedback, it sometimes makes mistakes, suggesting that human tutor scores are still necessary to balance out potential inaccuracies from the AI. Participant 6 expressed the valuable aspects of the tutor's scoring:

> The tutor's scoring is valuable as it provides a clear understanding of the current speaking level. Additionally, direct interaction allows for clarification if there are aspects that are not fully understood. This ensures the given score is transparent, with clear explanations of both the strengths and areas for improvement. (P6)

A key theme that emerges from the responses is the perception of objectivity. Two participants, argue that SpeechAce's scoring is more objective than that of human tutors, given that AI is not influenced by individual biases or inconsistencies. P7 appreciated the consistency and accuracy that SpeechAce offers, noting that human tutors may have variations in judgment due to their own limitations, such as hearing issues or differing evaluation criteria.

> But if it is about scoring, I think computer assistants such as SpeechAce would be more objective and accurate if its programming is done appropriately. I very much appreciate the opinions and scores from IELTS tutors, but as for official test scores, scoring done by programmed computers would be more fair for candidates. (P7)

In conclusion, while SpeechAce is valued for its efficiency, speed, and objective scoring, human tutors are still regarded as essential for their ability to provide more nuanced, comprehensive, and accurate assessments that are largely considered to be objective as well.

Learners' Perceptions on the Feedback

A consistent preference for a blended approach, integrating both AI-generated and human feedback, emerged from five participants. The majority (eight participants) recognized AI's strengths, specifically its swift, objective scoring and accurate transcription of speaking responses. However, they also noted that AI feedback could sometimes miss subtler aspects of their performance. As P1 put it,

> Human tutor's feedback provides more detailed feedback and concrete suggestions for improvement, such as replacing vocabulary with more appropriate idioms. This way, we immediately understand how to use them in the right context. (P1)

In spite of the statement given by three participants that AI excels in offering structured category-based feedback, such as highlighting vocabulary usage or identifying areas for improvement (P5, P6, P9), human feedback is seen as more valuable for its depth and personalization by the others. Eight participants emphasize that human tutors provide more specific, context-aware insights, often drawing from their own experiences to address individual challenges and offer practical suggestions. Feedback from human tutors is considered especially important for its ability to provide tailored advice on areas like pronunciation, emotional delivery, and overall communication style, aspects that AI may not fully capture. P2 described this perspective as follows:

> Of course, Tutor IELTS is better because they provide more comprehensive feedback on our speaking test. They go beyond just scoring and offer in-depth comments that specifically address our weaknesses. This allows us to gain a clearer understanding of the areas where we need improvement, whether it's in pronunciation, emotional delivery, overall communication style, fluency, grammar, or vocabulary. The feedback is tailored to our individual performance, which makes it much more valuable and effective in helping us improve our speaking skills for the test. (P2)

Overall, the majority of participants (seven participants) favor a combination of both AI and human feedback, seeing them as complementary. The objectivity and speed of AI tools, like SpeechAce, offer clear advantages for self-assessment, while human tutors were still better in providing the more detailed, individualized guidance necessary for comprehensive language learning. P9 expressed the desire to have both types of feedback:

> I really like both because using them together helps improve the speaking section more effectively. SpeechAce gives quick, clear feedback on areas like mispronunciations and fluency numbers, while Tutor IELTS offers advice on things like stresses and intonation. Combining both gives me a complete picture of where I can improve and helps me practice more efficiently. (P9)

These responses suggest the desirability of combining human and computer feedback rather than choosing just one over the other, because students see them working together to offer a complete approach to supporting their learning improvement together. Computer feedback, as

given by SpeechAce, is fast and perceivably accurate but lacks the personal and specific ideas for improvement. Human feedback, on the other hand, was seen as giving deeper insights into things like pronunciation features and practical steps for improvement.

Discussion

Overall, the data analysis shows that participants have different views on the use of AI tools like SpeechAce compared to human tutors for IELTS speaking preparation. Participants liked that SpeechAce offered flexibility and was easy to use. SpeechAce allowed learners to practice at any time and place, fitting their schedules and making them less nervous about speaking with a live examiner. They also valued the fast scoring and feedback from the AI tool for checking their progress and finding areas to improve. However, participants also emphasized how important human tutors are for giving feedback that is personal, detailed, and context sensitive. They pointed out that human tutors can give specific advice, address individual weaknesses, and evaluate speaking in a more complete way, including relevant factors like body language and how they interact. They valued the helpful learning atmosphere provided by the tutors and the way the tutors changed their feedback to fit each learner's needs. Considering these findings, the study analysis indicates that learners see strengths of both AI tools and human tutors, even though their strengths are different. Many participants prefer using both, combining the efficiency and objectivity of AI with the personal help and full feedback of human interaction.

These differences between the personalized feedback of the tutor and description-based feedback of the SpeechAce were distinguishable to participants despite the objective similarities in how both SpeechAce and human tutors provide feedback on speaking performance. Both rely on general descriptions that are somewhat standardized and based on established scoring rubrics like the IELTS speaking criteria. SpeechAce provides feedback by using objective data points, such as word count, the number and placement of pauses to assess fluency, and natural language processing (NLP) technology to evaluate whether pronunciation is correct. These feedback points are based on quantitative metrics, which focus on specific aspects of speech performance, such as fluency and pronunciation, but they do not take into account idiosyncratic pronunciation characteristics or other features like lexicogrammar and coherence, thus neglecting learners' individual speech patterns or errors.

The human tutor offers feedback that also references the IELTS speaking rubric but tends to focus on different aspects of the learner's speaking. Tutors typically point out individual mispronounced words, assess perceived fluency, and consider other features of pronunciation, such as intonation and stress, in addition to flagging lexicogrammatical choices that affect speaking performance. Human tutors provide personalized feedback based on the individual learner's unique speaking patterns even through the overall structure of the feedback often aligns with the more general descriptions used for the IELTS speaking. As a result, the types of feedback offered by the tutor and by SpeechAce on the surface may seem similar, especially when the feedback focuses on standardized speech features like fluency and pronunciation, which are commonly evaluated in both approaches. Nevertheless, participants perceived some important differences between the two.

It should be recognized however that the perceptions and experiences of the participants reported here may only be applicable to the participants involved in this study and that perceptions and experiences may vary across individuals, settings, and cultural backgrounds. Moreover, this speaking test simulation was all carried out remotely, instead of in-person. This condition may have affected participants' speaking performance and perceptions (Nakatsuhara et al., 2017, 2021). Future research is needed to investigate the generalizability of the results by comparing perceptions of learners across multiple contexts using quantitative research methods (O'Sullivan et al., 2023). A quantitative study can provide measurable, replicable results that are more easily compared across larger groups or settings. By employing statistical tools, researchers can analyze larger sample sizes, control for confounding variables, and test hypotheses in a more structured manner. One proposed idea is to use correlational and regression analyses to examine whether higher language proficiency in learners is associated with more positive of negative perceptions of SpeechAce. Correlational analysis would identify the strength and direction of the relationship between proficiency and perceptions, while regression analysis could predict how much proficiency accounts for variance in learners' views. By controlling for other factors, these methods could offer insights into how language proficiency impacts users' experiences with the tool, helping to identify one factor that may enhance or hinder its effectiveness for different proficiency levels.

Conclusion

This study shows that test simulations administered and scored by humans, with feedback from human examiners, were perceived by users to be different from those carried out by an AI tool. Results indicated that neither method can fully replace the other, meaning AI cannot completely replace human involvement. Results suggest that amplifying HAI interaction (Rasmussen & Andersen, 1991) with a combination of both human and AI feedback may offer the most satisfying method for language learners. By integrating the strengths of human expertise and AI efficiency, learners may experience the benefit of a richer, more balanced educational experience that enhances both the accuracy and personalization of their learning journey.

Building on the study's findings, future research could expand in several key directions. One direction would be to provide ample access to AI's structured and immediate feedback for beginners, while satisfying more advanced learners with the personalized in-depth assessment provided by human tutors. Research should analyze the success of such targeted use of AI tools to better meet the needs of learners at different proficiency levels. A second direction would be to explore the integration of a two-way communication AI powered humanoid avatar, such as a chatbot, for speaking test preparation Unlike the one-way AI tool used in this study, interactive AI systems may allow for simulating realistic conversations and providing learners with the opportunity to ask clarifying questions (Chukharev-Hudilainen & Ockey, 2021; Gokturk & Chukharev-Hudilainen, 2023; Gokturk & Chukharev, 2024). This technology may support greater learner engagement, enhance fluency, and better prepare learners for real-world speaking scenarios.

A third direction would be to conduct a longitudinal study to assess the long-term impact of combining AI and human feedback on language proficiency to provide deeper insights into the effectiveness of this hybrid approach over time. A well-structured longitudinal study could involve testing participants at multiple time intervals, such as 1 month, 3 months, and 6 months, using the AI platform alongside human-administered test simulations. By including a control group that only uses human-administered test simulations, researchers can compare the outcomes of the hybrid approach to traditional methods. This type of study would allow for a more comprehensive understanding of how the combination of AI and human feedback influences language learning progression, retention, and overall improvement in language proficiency over time, providing valuable data on the sustained benefits (or potential drawbacks) of integrating both methods. These next steps would refine AI tools, provide more personalized learning, and offer deeper insights into the dynamics of language acquisition through technology.

References

Celce-Murcia, M., Brinton, D. M., & Snow, M. A. (Eds.). (2014). *Teaching English as a second or foreign language* (4th ed.). Heinle

Chapelle, C. A. (2013). Reliability in language assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 4918-4923). Blackwell/Wiley. https://doi.org/10.1002/9781405198431.wbeal1003

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment. *Language Learning & Technology, 20*(2), 116–128. https://doi.org/10.64152/10125/44464

Chukharev-Hudilainen, E., & Ockey, G. J. (2021). *The development and evaluation of interactional competence elicitor for oral language assessments (ETS Research Report Series vol. 2021 issue 1)*. Educational Testing Service. https://doi.org/10.1002/ets2.12319

Gokturk, N., & Chukharev-Hudilainen, E. (2023). Strategy use in a spoken dialog system-delivered paired discussion task: A stimulated recall study. *Language Testing, 40*(3), 630-657. https://doi.org/10.1177/02655322231152620

Gokturk, N., & Chukharev, E. (2024). Exploring the potential of a spoken dialog system-delivered paired discussion task for assessing interactional competence. *Language Assessment Quarterly, 21*(1), 60–99. https://doi.org/10.1080/15434303.2023.2289173

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 279–294). Routledge. https://doi.org/10.4324/9781315145006

Isotalus, P., Eklund, M., & Karppinen, K. (2024). Artificial intelligence as a feedback provider in practicing public speaking. *Communication Teacher, 39*(1), 78–85. https://doi.org/10.1080/17404622.2024.2407910

Khabbazbashi, N., Xu, J., & Galaczi, E. (2021). Opening the black box: Exploring automated speaking evaluation. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds.), *Issues in language testing around the world: Insights for language test users* (pp. 333–343). Springer. https://doi.org/10.1007/978-981-33-4232-3_25

Loukina, A., & Buzick, H. (2017). *Use of automated scoring in spoken language assessments for test takers with speech impairments* (*Research Report ETS RR–17-42*). Educational Testing Service. https://doi.org/10.1002/ets2.12170

Macinska, S. & Vinkler, A. (2024) The role of AI in language learning: Research evidence and strategies for success. *Cambridge Papers in English Language Education.* Cambridge University Press. https://perma.cc/D47Q-HFXC

Martin, F., Stamper, B., & Flowers, C. (2020). Examining student perception of their readiness for online learning: Importance and confidence. *Online Learning, 24*(2), 38-58. https://doi.org/10.24059/olj.v24i2.2053

McNamara, T. (2019). *Fairness, justice and language assessment*. Oxford University Press.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly, 14*(1), 1-18. https://doi.org/10.1080/15434303.2016.1263637

Nakatsuhara, F., Khabbazbashi, N., & Inoue, C. (2021). Assessing speaking. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 209-222). Routledge. https://doi.org/10.4324/9781003220756

O'Sullivan, B., Breakspear, T., & Bayliss, W. (2023). Validating an AI-driven scoring system: The model card approach. In K. Sadeghi & D. Douglas (Eds.), *Fundamental considerations in technology-mediated language assessment* (pp. 115–134). Routledge. https://doi.org/10.4324/9781003292395-10

Pataranutaporn, P., Danry, V., Leong, J., Novy, D., Maes, P. & Shra, M. (2021) AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, *3*, 1013–1022. https://doi.org/10.1038/s42256-021-00417-9

Pearson. (2023, August 8). *Versant English language test: Assessment tools for business*. https://perma.cc/4V7M-SJHB

Rasmussen, J., & Andersen, H. B. (Eds.). (1991). *Human-computer interaction. Research directions in cognitive science.* Lawrence Erlbaum Associates.

Read, J. (2022). Test review: The International English Language Testing System (IELTS). *Language Testing, 39*(4), 679-694. https://doi.org/10.1177/02655322221086211

Sadeghi, K., & Douglas, D. (Eds.). (2023). *Fundamental considerations in technology-mediated language assessment*. Routledge. https://doi.org/10.4324/9781003292395

Schiefelbein, J. S. (2023). *Human vs machine: Hyper-realistic avatars and their efficacy as a communication channel* [Unpublished master's thesis]. University of South Florida. https://digitalcommons.usf.edu/etd/10088

Schunk, D.H., & Meece, J.L. (Eds.). (1992). *Student perceptions in the classroom* (1st ed.). Routledge. https://doi.org/10.4324/9780203052532

Sharples, M., & Pérez y Pérez, R. (2022). *Story machines: How computers have become creative writers* (1st ed.). Routledge. https://doi.org/10.4324/9781003161431

Susanto, A. (2018). Revisiting the concept of guru in various cultural perspectives. *Magister Scientiae, 44,* 239-256. https://doi.org/10.33508/mgs.v2i44.1977

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

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341-358). De Gruyter Mouton. https://doi.org/10.1515/9781614513827-023

Woo, J. H., & Choi, H. Y. (2021). Systematic review for AI-based language learning tools. *Journal of Digital Contents Society, 22*(11), 1783-1792. https://doi.org/10.9728/dcs.2021.22.11.1783

Xi, X., Mollaun, P., & Davis, L. (2024). Use of formulaic language in TOEFL iBT Speaking responses: Implications for scoring. In L. Davis & J. M. Norris (Eds.), *Challenges and innovations in speaking assessment* (1st ed., pp. 125-145). Routledge. https://doi.org/10.4324/9781315201504

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X., & Yoon, S. (2015). *Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFTTM)* (*Research Report ETS RR–15-31)*. Educational Testing Service. https://doi.org/10.1002/ets2.12080

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech* (1st ed.). Routledge. https://doi.org/10.4324/9781315165103

Zhu, Q. (2023). Empowering language learning through IoT and big data: An innovative English translation approach. *Soft Computing, 27*(17), 12725–12740. https://doi.org/10.1007/s00500-023-08850-1

About the Author

Andrias Susanto is a Ph.D. student at Iowa State University, specializing in Applied Linguistics and Technology as well as Human-Computer Interaction. His research focuses on technology-enhanced language assessment, with particular interest in mixed-reality environments and artificial intelligence applications to foster self-regulated learning. He has presented his work at leading conferences such as AsiaTEFL, TSLL, PSLLT, CALICO, and AAAL. His work has also been published in peer-reviewed journals including *Language Learning & Technology*, *CALICO Journal*, *International Journal of Applied Linguistics*, and *TESOL Journal*, among others.

Chapter 9

Comparing Human Raters' and ChatGPT-4's Scores in Measuring Content in CLIL Presentation

Rie Koizumi[1], Saki Suemori[1], and Yusuke Kubo[2]
[1]*University of Tsukuba*
[2]*University of Tsukuba/University of Yamanashi*

Considering the improvement of AI in accessibility and response quality, this study examines the degree to which ChatGPT-4 generates scores comparable to human raters on content aspects in Content and Language Integrated Learning (CLIL) presentations, where both content and language instruction are prioritized, necessitating effective content assessment. We obtained 52 presentation scripts from 37 university students who had taken a CLIL course. Using a detailed prompt that includes an analytic rubric, its explanation, example scripts, and benchmarking scores, we asked ChatGPT-4 to score each presentation script. This procedure was repeated three times, and the scores were compared with human-rated scores primarily using many-facet Rasch measurement. The results showed that seven human raters and ChatGPT-4 produced consistent scores, with ChatGPT-4 tending to produce slightly lenient ratings and fewer biases. We discuss the implications of using ChatGPT-4 as a supplementary rater.

Introduction

Following advancements in artificial intelligence (AI) accessibility and response quality, interest in automated scoring for learners' responses has grown. Previous research has compared scores between human raters and the Generative AI tool, ChatGPT. For example, regarding scoring second language (L2) writing, Mizumoto and Eguchi (2023) compared ChatGPT-3.5 with human raters, finding that it yielded relatively comparable holistic scores, with exact and adjacent-level agreements of 54.33% and 89.15%, respectively. Similarly, Kim et al. (2024) reported some agreement between ChatGPT-4 ratings and human raters regarding scoring L2 integrated

Cite as: Koizumi, R., Suemori, S., & Kubo, Y. (2025). Comparing human raters' and ChatGPT-4's scores in measuring content in CLIL presentation. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 166-196). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.09
©2025 Koizumi, Suemori, and Kubo. Published under a CC BY license.

reading-to-writing essays, with an exact agreement of 57% with writing prompt and source texts, and 55% without them. Additionally, Yamashita (2024) examined the comparability of ChatGPT-4 and human ratings of L2 essays using an analytic rubric of complexity, accuracy, and fluency to score L2 writing. Employing many-facet Rasch measurement (MFRM), he demonstrated that ChatGPT-4's mean scores were close to the means of human scores; however, ChatGPT-4 generated much less variation across L2 writers than human raters did, with the AI yielding predictable scores (infit mean squares [MSs] of 0.23). In a similar vein, Shin and Lee (2024) used an analytic rubric of task completion, content, organization, and language use, employing MFRM, and showed that ChatGPT-4 (using My GPTs of ChatGPT Plus) yielded slightly more lenient scores compared to human raters and showed biases associated with the criteria. The standard deviations were slightly smaller, but similar (e.g., 1.07 for ChatGPT vs. 1.44–1.47 for human-raters). ChatGPT-3 ratings also fit the model, showing scoring consistency (infit MS = 1.33).

In other words, previous studies suggest moderate agreement between ChatGPT and human ratings. Although there may be concerns regarding the differences between ChatGPT and human ratings, disparate ratings can also be observed among trained raters (Knoch et al., 2021). Unlike human ratings however, one persistent concern is the limited score variations in the ratings produced by ChatGPT, a tendency that is different from that of humans. For instance, in Yamashita's (2024) study, ratings from ChatGPT displayed smaller variations, and Mizumoto and Eguchi (2023) reported that ChatGPT failed to use the lowest and highest levels of the holistic scale, indicating use of a slightly narrow rubric range. Both studies used zero-shot prompts without sample essays or benchmark scores. This lack of detailed information provided to ChatGPT for scoring may have contributed to the resulting limited score variations. This speculation was supported by Shin and Lee (2024), who used a detailed prompt with additional contextual information, such as benchmarking scripts and the scores for each level. Also, Xu et al. (2024) reported an improvement in the correlation between ChatGPT-4 and human-identified errors when a more detailed prompt was used to identify grammatical errors in L2 writing (see also Kim et al., 2024). Accordingly, the current study employed a detailed prompt to produce similar score variations between ChatGPT's and humans' assessment of the content of test takers' responses.

Previous studies have predominantly focused on the linguistic aspects of L2 proficiency, such as accuracy and fluency (Yamashita, 2024) or general proficiency, as evaluated by holistic rubrics (e.g., Mizumoto & Eguchi, 2023). Others have evaluated general rather than academic content (e.g., Shin & Lee, 2024). However, evaluating linguistic features is generally simpler than assessing content features, as the former may be often more straightforward to identify at the surface level. Consequently, assessing content aspects poses significant challenges for humans (Sato, 2024) and possibly AI scoring systems, and these have not yet been well examined.

Despite the challenges, evaluating content is indispensable in L2 teaching, particularly in Content and Language Integrated Learning (CLIL) contexts. Because teachers focus on both content and language learning in class, they are expected to evaluate both aspects. However,

assessing CLIL content is challenging for L2 teachers because assessment methods are not well researched (deBoer & Leontjev, 2020).

To clarify the content construct in CLIL contexts, Sato (2024) analyzed subject specialists' perspectives on the content quality of essays completed by L2 learners in a CLIL course. The results showed that they mainly examined two types of content features: (a) content features in general L2 writing (clarity, logicality, elaboration, complexity, coverage, and discourse structure and organization) and (b) content specific to L2 writing in a CLIL context (content accuracy, and citation and research).

Using the content aspects proposed by Sato (2024), this study reports the development of a content-specific rubric and its application by human raters and ChatGPT to establish an effective method for evaluating content. This study aims to explore the potential of ChatGPT-4 as a supplementary tool in the assessment process rather than a replacement of human raters. This approach aligns with the perspective outlined by Xi (2023), who proposed acceptable use cases for AI under black-box conditions: (a) when the tests are low stakes; (b) when AI is used for confirmation or supplementary ratings; (c) when stakeholders do not require transparency; (d) when the content and expressions are constrained; and (e) when AI is employed for voice recognition, feature extraction, and computation rather than as a standalone scoring model. In the context of CLIL assessment, conditions (a) and (b) are particularly relevant. However, further investigations are required to ensure scoring precision when human raters and ChatGPT ratings are used complementarily.

To address the need to develop a content rubric to facilitate the assessment of content in CLIL classrooms and examine the feasibility of scoring content by AI, this study examines the degree to which ChatGPT-4 can yield scores comparable to those produced by human raters for scoring content aspects in CLIL presentation scripts. For this purpose, we examine agreement (e.g., Kim et al., 2024; Mizumoto & Eguchi, 2023), severity, consistency, and biases (Shin & Lee, 2024; Yamashita, 2024) between human and ChatGPT ratings. We pose three research questions (RQs) to evaluate L2 English learners' presentation scripts in CLIL classrooms:

1. To what degree do ChatGPT-4 ratings agree with human ratings?
2. How do ChatGPT-4 ratings compare with human ratings in terms of severity and consistency?
3. To what degree does ChatGPT-4 exhibit biases toward test takers, tasks, and criteria?

Methods

Participants and Contexts

Thirty-seven students from a private university in Japan submitted presentation scripts. Most students (86.49%) had a CEFR A2 level as per their scores of the TOEIC Listening and Reading test (Test of English for International Communication; https://www.ets.org/pdfs/toeic/toeic-mapping-cefr-reference.pdf). They had studied English as a foreign language in secondary school for at least six years, and took a one-term CLIL course in English, with an introduction to

L2 acquisition and assessment module. They were required to make presentations and submit their scripts and slides through an online system in advance.

Tasks

We collected 52 scripts (see Table 9.1). While the presentation scripts were aimed at speaking, they were considered as writing tasks because they were written. Group 1 submitted Task 1 scripts as their midterm assignment, and Groups 1 and 2 submitted Tasks 2 and 3 as their final assignment, respectively. Both groups belonged to intact classes and took the CLIL courses that the first author taught in spring and fall terms, respectively. Students were informed that their presentations would be evaluated using a rubric with four criteria: content, use of technical terms learned in the courses (e.g., input/output), delivery, and language; these were different from the rubric developed in this study. We assessed only the scripts and slides, rather than their oral presentations (see Supplementary Materials).

Table 9.1

Participants and Tasks

Participant	Task	Task and topic	Main Prompt
Group 1 (*n* = 18)	Task 1 (*n* = 15)	L2 acquisition	Talk about your English (or another foreign language) learning experiences. Were they effective? Why or why not? How can you improve your L2 learning?
	Task 2 (*n* = 18)	L2 assessment	Tests can be a tool to encourage you to study hard. What test would you like to study to increase your speaking ability? Why? Describe the test in detail.
Group 2 (*n* = 19)	Task 3	L2 acquisition	Talk about a good material that helps you improve your English and that you want to recommend to your classmates.

Note. We included one midterm and two final tasks. To examine the study focus more comprehensively, future studies should include one more midterm task to strike a balance between midterm and final assignments.

Rubric

Based on Sato (2024), a content rubric was developed using five criteria: task achievement, clarity of content, elaboration, logicality, and content accuracy, scored from 0 to 4 (see Figure 9.1). The categories were ordered so raters found them easy to evaluate and did not assume importance of any category.

Figure 9.1

Rubric for Assessing Content in CLIL Classroom

	Task Achievement	Clarity	Elaboration	Logicality	Accuracy
4	Provides a sufficient response to the writing prompt and writes extensively on the topic.	The content (including message and meaning) is sufficiently clear.	There are explanations, examples, and reasons, and they are sufficiently detailed (with sufficient quantity).	The flow of ideas or paragraphs is sufficiently logical, and paragraphs are sufficiently linked.	The ideas, information, and premises are sufficiently accurate and comprehensive.
3	Provides a mostly sufficient response to the prompt and writes largely on the topic.	The content is moderately clear.	There are explanations, examples, and reasons, and the level of detail is moderately high.	The flow of ideas or paragraphs is moderately logical, and paragraphs are fairly linked.	The ideas, information, and premises are moderately accurate and comprehensive.
2	Provides a partial response to the prompt and writes partially on the topic.	The content is partially clear.	There are explanations, examples, and reasons, but they are only partially detailed.	The flow of ideas or paragraphs is only partially logical, and paragraphs are only partially linked.	The ideas, information, and premises are only partially accurate and comprehensive.
1	Does not respond to the prompt and does not write on the topic.	The content is unclear.	There are no explanations, examples, or reasons.	The flow of ideas is not logical, and paragraphs are not linked.	The ideas, information, and premises are not accurate or comprehensive.
0	Too short to assess.	Too short to assess.	Too short to assess.	Too short to assess.	Too short to assess.

We set Level 2 in the rubric as the minimum passing level, Levels 3 and 4 as "mostly sufficient" and "sufficient", Level 1 as failing to provide required contents, and Level 0 as too short a script. After creating the rubric, we selected 12 representative scripts and determined benchmark scores, providing rationales for them.

Human Raters

Table 9.2 shows each rater's characteristics. Raters A to C developed the rubric and discussed benchmarking scores and rationales. The other raters were invited to evaluate scripts and received an honorarium after the completion. We included raters with diverse backgrounds to examine the rubric's applicability. Most raters did not know the test takers, except for Rater A, who had taught them in the CLIL classroom; however, scoring was conducted more than one year later than the actual teaching and she remembered little about the students.

Seven human raters underwent online training for at least three hours, during which they evaluated the scripts, discussed disagreements, and revised rationales for benchmarking scores. During scoring, raters would only read scripts, except when students clearly referred to slides, such as "Here is the outline of my presentation." When raters were unsure about the level, they were asked to score strictly. After the training, they independently evaluated the scripts and provided scores for each student, task, and criterion. When they were unsure or encountered problems in scoring, they reported their scoring uncertainty.

Table 9.2

Characteristics of Human Raters

Rater	Position	Knowledge about content	Teaching experience	Scoring experience	Task rated
A	University teacher	✓	✓	✓	3, 2, 1
B	University teacher	✓	✓		1, 3
C	Graduate student	✓	✓	✓	2, 1
D	Former junior high school teacher		✓	✓	1, 2, 3
E to G	Undergraduate student	Partially ✓			1, 2, 3 2, 3, 1 3, 1, 2

Note. Raters A and B had the CEFR C1 level; Raters C to G had CEFR B2 levels (estimated from Eiken Grade certificates and https://www.eiken.or.jp/eiken/en/grades/). Raters E to G were studying to receive teaching certifications at secondary school.

✓ = Yes. Partially ✓ = They learned introductory contents at the undergraduate courses. In the column "Task rated," 1, 2, 3 reflects the order of scoring: Tasks 1, 2, and 3. As they were considered raters with content knowledge, Raters A, B, and C evaluated Tasks 1 to 3, Tasks 1 and 3, and Tasks 1 and 2, respectively, according to their specialty (i.e., L2 acquisition and/or L2 assessment). Rater F did not finish scoring due to the limited time (see Table 9.4 for the number of ratings that each rater provided).

ChatGPT-4 Ratings

To effectively use ChatGPT-4, we created a detailed prompt for each task with a description of the tasks and scripts to be scored, example scripts, and benchmark scores, as shown in Figure 9.2. Although the prompt content was similar to the training materials the raters viewed during their training, the prompt did not have specific explanations for why a certain score was reached to avoid making the prompt long and complicated. Furthermore, while human raters were allowed to read slides when necessary, ChatGPT-4 was not (ChatGPT-4 is expressed as "ChatGPT" hereafter).

Figure 9.2

Detailed Prompt Structure of Training for Scoring Task 1 for ChatGPT

Role & Task	I would like you to mark a presentation script written by Japanese learners of English as a foreign language (EFL) learners. Each script is scored on a scale of 0 to 4, with 4 being the highest and 0 the lowest. You need to give me the reason why you assign that specific score. Report both the scores and the reasons.
Explanation of the data	The directions for the students are as follows: Write a presentation script. You have 2 min (between 1 min and 3 min) to make a presentation. *(omitted)* Points to include: Introduction Topic 1: Talk about your English (or another foreign language) learning experiences. Were they effective? Why or why not? Topic 2: How can you improve your L2 (second language) learning? Conclusion: Summarize your talk briefly.
Rubric	The script is scored based on the following rubric: Task Achievement: The extent to which the script answers the question and is relevant to the topic. [score 4] - Provides a sufficient response to the writing prompt and writes extensively on the topic. *(omitted)* Note: 4 = A; 3 = B+; 2 = B; 1 = C. Generally, a score of 2 = B is considered a passing level. Difference between 3 and 2: A score of 3 indicates that there are some (not many) areas where the criteria are not met, while a score of 2 indicates that there are many areas where the criteria are not met.
Sample texts and benchmarking scores	Below is the data for you to familiarize yourself with scoring using the topic and the rubric above: Training 1: Data 1: No. X *(Actual script)* This response is scored 1, 1, 0, 0, and 0 for Task Achievement, Clarity of Content, Elaboration, Logicality, and Accuracy of Content, respectively. *(omitted)*
Specific instructions for ChatGPT	Now, let' s start to score the following script. There are five criteria: Task Achievement, Clarity of Content, Elaboration, Logicality, and Accuracy of Content. Please give a score for each criterion and give me the reason why you assign that specific score. Report both the scores and the reasons. Please save the scores and reasons for each script in Microsoft Excel format and export it. Here is the template for the output. *(omitted)*

Note. Data = Task. See Appendix A for the whole prompt.

A research assistant pasted the prompt and each script individually (from Scripts 1 to 52, in the order of Tasks 1, 2, and 3) in a ChatGPT chat box in the web browser, to generate rubric-based scores (k = 52). She repeated this procedure three times between February 10 and

15, 2024 and noted problems with the ChatGPT outputs; she conducted the first- to third-time trials on different days. All 52 scripts were processed in the same chat for the first trial, and another chat was created for the second and third trials each; this procedure was undertaken to simulate human raters' scoring. When ChatGPT did not provide the requested information, the same request was inputted to obtain the scores and reasons.

Analysis

To examine the relationship between human and ChatGPT ratings, we created four datasets (Table 9.3). Dataset 0 was the baseline (without ChatGPT ratings), and Datasets 1 to 3 were used to observe how rating results changed owing to the addition of the first, second, or third ChatGPT trials. We did not analyze ratings of the first to third ChatGPT trials together because they were dependent, produced using the same MFRM mechanism; MRFM assumes that raters are independent (DeMars, 2021; see details on MFRM below).

Table 9.3

Four Datasets Analyzed

Dataset	No. of Raters
0 (Humans)	7
1 (Humans + the first trial of ChatGPT [ChatGPT 1])	8
2 (Humans + the second trial of ChatGPT [ChatGPT 2])	8
3 (Humans + the third trial of ChatGPT [ChatGPT 3])	8

For the first RQ, we primarily used Fleiss' kappa to examine multiple raters' agreement with ordinal data, using the LangTest interrater agreement application (https://langtest.jp/shiny/kappa/). The second and third RQs were examined using MFRM with a rating scale model and bias analysis, employing Facets (Ver. 4.2.3; Linacre, 2024a) with four facets (student x task x rater x criterion).

Bond et al. (2021) and McNamara et al. (2019) state that MFRM is an extension of basic Rasch measurement, the latter of which basic Rasch measurement encompasses a family of probabilistic models that predict the performance of test takers and tasks; MFRM extends the basic measurement by including additional facets, such as raters and criteria. When two or more criteria are present in the rubric, two types of models are available: rating scale and partial credit. The rating scale model assumes the same structure across criteria (e.g., the same difficulty distance between Levels 1 and 2 across content clarity and logicality in the rubric); the partial credit model does not assume such a structure, estimates the structures for each criterion, and requires more data than the rating scale model due to more complex estimation. We used both, and report results of the rating scale model, owing to the limited number of participants; however, the results between the two models were comparable (see Supplementary Materials for the results using a partial credit model).

We also examined model fit to determine what extent performance patterns of test takers, tasks, raters, and criteria align with those that the Rasch model predicts; two types of

misfit exist: underfit when data patterns are too distant from the predicted patterns and overfit when too close (Bond et al., 2021; McNamara et al., 2019). Owing to the low-stakes nature of the current assessment, we considered infit MSs between 0.5 and 1.5 as acceptable, with infit MSs more than 1.5 and less than 0.5 as underfit and overfit, respectively (Linacre, 2024b). Bias was detected when *p*-values were less than .05, as determined by *t*-statistics (Linacre, 2024b). While missing responses were present in the data, MFRM could handle such missing data (see Linacre, 2024b); because scripts were scored by at least six out of seven raters, comparability regarding ratings across tasks and groups was not considered problematic.

Results

Before reporting the detailed results addressing each research question, we report the descriptive statistics for each human and ChatGPT rating (see Table 9.4). The mean scores on the 0–4 scale were generally similar, but those of ChatGPTs 1–3 were slightly higher.

Table 9.4

Descriptive Statistics of All the Criteria Scores Combined

Rater	No. of ratings	Mean	*SD*	Min	Max
A	260	2.42	1.17	0	4
B	195	2.22	1.07	0	4
C	185	2.28	1.18	0	4
D	260	2.15	1.09	0	4
E	260	2.13	1.13	0	4
F	195	2.29	1.11	0	4
G	260	2.30	1.05	0	4
ChatGPT 1	260	2.64	1.13	0	4
ChatGPT 2	260	2.58	1.15	0	4
ChatGPT 3	260	2.65	1.13	0	4

Rater Agreement Across Criteria

To answer research question 1, we used Raters A, D, and G, and ChatGPTs 1–3 to examine rater agreement (see Table 9.5). Three raters were selected from among raters who scored all scripts for all tasks to set the same number as ChatGPT's trials. Results showed that rater agreements were acceptable, but not excellent across criteria (e.g., $k = .46$ among the three humans in task achievement). The three ChatGPT ratings constantly yielded higher agreement than the three human ratings (e.g., .71 vs. .46 in task achievement). Among the five criteria, content accuracy had the lowest agreement, particularly among both human and ChatGPT ratings. The difficulty in scoring content accuracy was probably because this criterion typically requires expert content knowledge to determine the accuracy and representativeness of the content as indicated in the rubric in Figure 9.1.

Table 9.5

Rater Agreement (Fleiss' Kappa [95% Confidence Interval])

	3 humans and 3 ChatGPTs	3 humans	3 ChatGPTs
Task achievement	.49 [.45, .53]	.46 [.37, .55]	.71 [.61, .81]
Clarity	.46 [.41, .50]	.42 [.33, .51]	.71 [.61, .81]
Elaboration	.55 [.51, .60]	.56 [.47, .65]	.60 [.51, .69]
Logicality	.47 [.43, .51]	.48 [.38, .57]	.69 [.59, .78]
Content accuracy	.43 [.39, .47]	.51 [.42, .60]	.54 [.45, .63]

MFRM

To address the second and third research questions, we examined the global model fit to show how the data fit the MFRM model. Specifically, we evaluated the global model fit by calculating the proportions of standardized residuals (referred to as "unexpected responses" in the output) falling outside the thresholds of ±2 and ±3, with acceptable limits set at approximately 5% and 1%, respectively (Linacre, 2024b, p. 183). Table 9.6 (*Global fit*) indicates that the percentages of residuals exceeding these thresholds were 3.1–3.3% and 0.1%, demonstrating a satisfactory global model fit to the data. To examine unidimensionality, we examined (a) infit and outfit mean squares and (b) point-biserial correlations based on McNamara et al. (2019). Both measures, (a) and (b), displayed positive results for tasks, raters, criteria, and most of the test takers across datasets (e.g., (b) .24 or more, .40 or more, and .44 or more with tasks, raters, and criteria). These results suggest adequate unidimensionality was found.

Table 9.6

Global Model Fit and Facet Results

Dataset	0 (Humans)	1 (Humans +ChatGPT 1)	2 (Humans +ChatGPT 2)	3 (Humans +ChatGPT 3)
Global fit: Standardized residuals				
Outside ±3	0.1% (1/1615)	0.1% (2/1875)	0.1% (1/1875)	0.1% (1/1875)
Outside ±2	3.3% (53/1615)	3.1% (59/1875)	3.3% (62/1875)	3.3% (61/1875)
Student				
Logit *M*	0.38	0.48	0.45	0.47
(SE, *SD*)	(0.21, 1.01)	(0.20, 1.07)	(0.20, 1.04)	(0.20, 1.04)
Infit MSs of more than 1.50	5.41% (2/37)	5.41% (2/37)	5.41% (2/37)[b]	5.41% (2/37)[b]
Infit MSs of less than 0.50	10.81% (4/37)	8.11% (3/37)	8.11% (3/37)	8.11% (3/37)
Separation, Strata, Reliability	4.48, 6.31, .95	5.02, 7.03, .96	4.91, 6.89, .96	4.91, 6.88, .96
Task[a]				
Logit *M*	0.00	0.00	0.00	0.00
(SE, *SD*)	(0.06, 0.55)	(0.05, 0.53)	(0.05, 0.54)	(0.05, 0.55)
Separation, Strata, Reliability	9.77, 13.36, .99	10.02, 13.69, .99	10.31, 14.08, .99	10.40, 14.21, .99
Rater[a]				
Logit *M*	0.00	0.00	0.00	0.00
(SE, *SD*)	(0.09, 0.21)	(0.09, 0.30)	(0.09, 0.26)	(0.09, 0.30)
Separation, Strata, Reliability	2.16, 3.21, .82	3.28, 4.71, .92	2.89, 4.19, .89	3.33, 4.77, .92

Table 9.6 (continued)

Global Model Fit and Facet Results

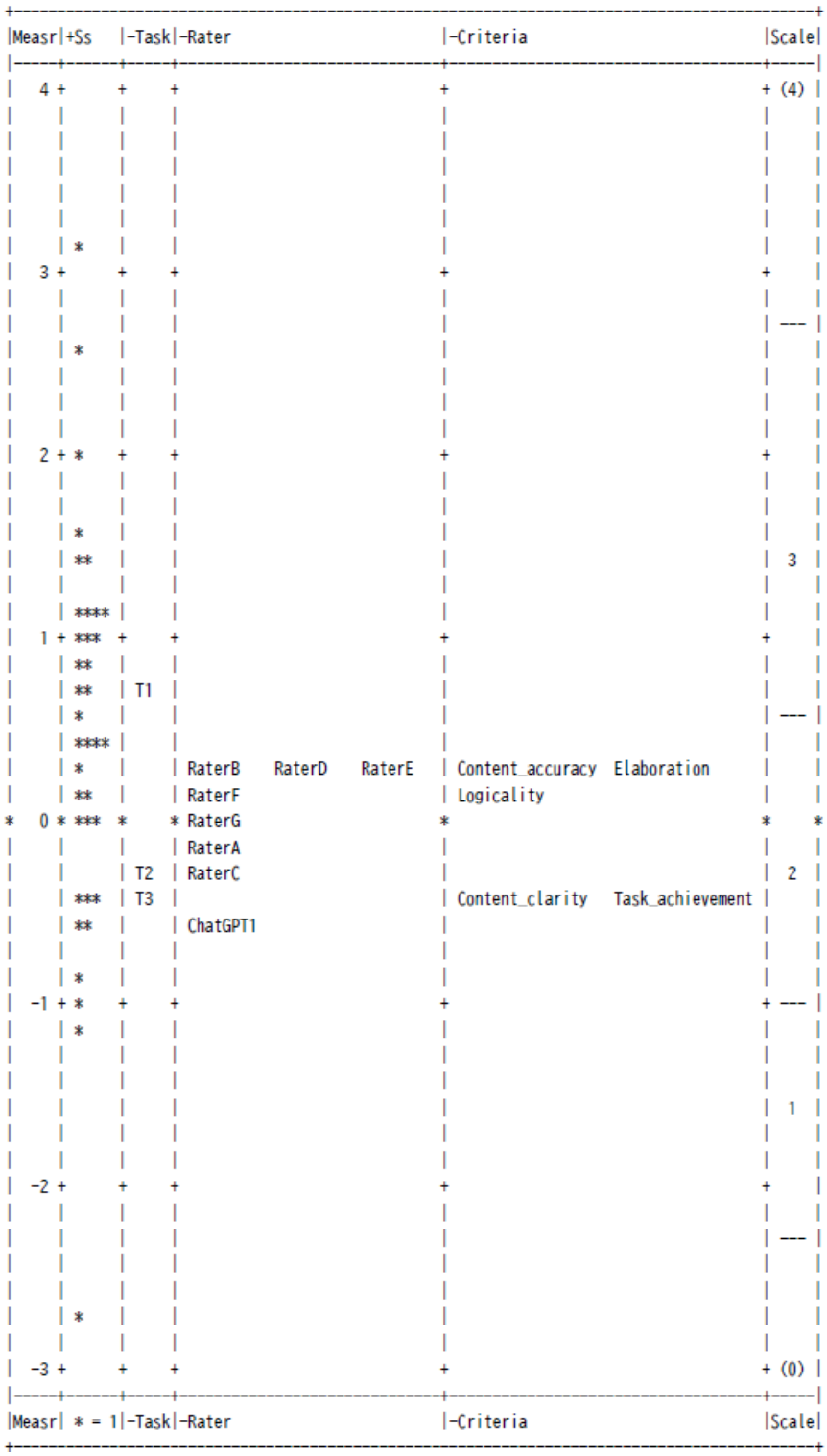
Dataset	0 (Humans)	1 (Humans +ChatGPT 1)	2 (Humans +ChatGPT 2)	3 (Humans +ChatGPT 3)
Exact vs. expected agreement	66.0% > 38.3%	63.5% > 38.1%	64.4% > 38.1%	64.0% > 38.0%
Criterion[a]				
Logit *M* (SE, *SD*)	0.00 (0.07, 0.32)	0.00 (0.07, 0.32)	0.00 (0.07, 0.31)	0.00 (0.07, 0.32)
Separation, Strata, Reliability	4.31, 6.08, .95	4.53, 6.38, .95	4.55, 6.39, .95	4.65, 6.53, .96
Criterion category statistics				
Satisfied criteria	3/5	3/5	3/5	3/5

Note. SE = Standard error. MSs = Mean squares. See Table 9.3 for an explanation of Datasets 0 to 3 (e.g., Humans+ChatGPT3). [a]No elements in each facet were outside the range of infit MS of 0.5 and 1.5. [b]One student who had infit MS of 2.1 misunderstood Task 1 instructions but successfully performed Task 2.

Figure 9.3 shows the relationships between facets in Dataset 1 with human raters and ChatGPT 1, for example, with 0 indicating the mean on the logit scale, and positive values indicating more able students, more difficult tasks and criteria, and more severe raters. The results suggested that the students' ability range was significantly wider than the variations of other facets. Task 1 was more difficult than Tasks 2 and 3, probably owing to both task difficulty and limited time and amount of work on the task during the lesson time; additionally, Task 1 was a midterm assignment, and students were not used to making presentation scripts based on what they learned in class and may have struggled to achieve Task 1 more than the other tasks. Raters B, D, and E were more severe, while ChatGPT 1 was the most lenient. However, all the raters were estimated to be within –1 and 1, indicating acceptable differences in rater severity. Among these criteria, content accuracy and elaboration were more difficult, whereas content clarity and task achievement were easier. This tendency was expected owing to the knowledge and skills required to express academic content and provide details on an academic topic. The scale on the right shows that average students (0.48 logits) received Levels 2 or 3 at 50% probability. The overall pictures were similar for Datasets 0, 2, and 3.

Figure 9.3

Variable Map for Dataset 1 (Humans + ChatGPT 1)



Note. Ratings of the first to third ChatGPT trials were analyzed separately to avoid dependency.

We further comprehensively examined the results for each facet across the datasets (Table 9.6, *Student*, *Task*, and *Criterion*). Generally, students' ability estimates were basically similar across datasets in means and *SD*s, (*M* = 0.38–0.48; *SD* = 1.01–1.07). Slight differences between humans and ChatGPTs were observed owing to the addition of lenient ChatGPT scores, which slightly increased the mean ability estimates (0.38 to 0.45–0.48). Students' scores fit the model well, with most having infit MSs between 0.50–1.50 (83.78–86.49%). Separation (4.48–5.02), strata (6.31–7.03), and reliability (.95–.96) were also similar, with students separated into four to seven ability groups, with high person reliability (which is conceptually the same as Cronbach's alpha). However, slight improvement in separation and strata were also observed owing to the addition of one more rater (ChatGPT; 4.48 to 5.02). These results aligned with those of the tasks and criteria. The raters' results are described below.

Results of criteria in terms of category statistics (see Table 9.6, *Criterion category statistics*) were also similar across datasets, based on Bond et al.'s (2021) five criteria (see Figure 9.4 for results of Dataset 1): First, each level of criteria (from 0–4) had more than 10 data counts (e.g., 173–624; see "Category counts, Used"). Second, average measures increased as the levels increased (e.g., –2.08–1.99 in "Quality control"). Threshold measures also increased along with level increase (–1.50–2.35 in "Rasch-Andrich"). Third, the level fit statistics were less than 2.0 (e.g., 0.5–1.2 in "Outfit MnSq [MSs]" in "quality control"). These results were positive, whereas the following two were negative. Fourth, the distance between Rasch-Andrich threshold measures should be between 1.4 and 5.0, but all the datasets showed acceptable distances except for a short distance between Levels 1 and 2 (i.e., 0.08–0.14; e.g., 0.14 = distance between –1.50 and –1.36). Fifth, the probability curve of the rubric (see Figure 9.5) shows the relationship between student ability estimates and the probability of obtaining a certain level; when ability estimates increase, students are likely to obtain higher levels. Although the probability curb should show a clear peak at each level, Level 1 did not, indicating that Level 1 was not clearly separated from the neighboring levels. These results suggest the need for modification of the levels or training materials. Because Level 2 is the passing level and there is a meaning between Levels 1 and 2, it is unlikely that we shall combine Levels 1 and 2. When we preliminarily combined Levels 0 and 1 and created a new Level 1, we obtained better category statistics, which may be a better move for rubric revision (see Appendix B).

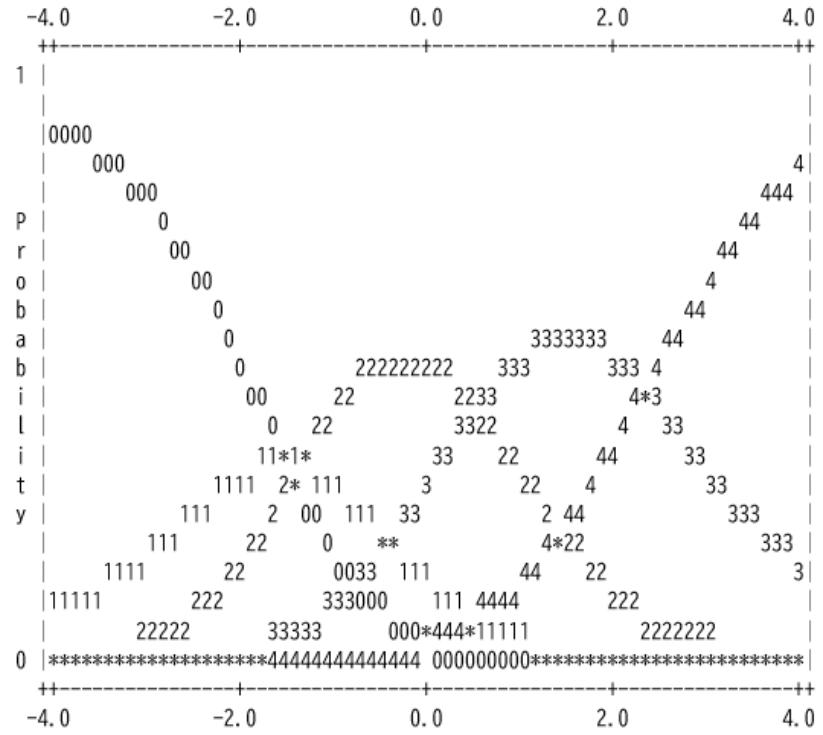
Figure 9.4

Category Statistics of the Rubric (Dataset 1: Humans + ChatGPT 1)

DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
0	173	173	9%	9%	-2.08	-1.83	.5			(-2.97)		low	low	100%
1	217	217	12%	21%	-.71	-.78	1.2	-1.50	.10	-1.55	-2.31	-1.50	-1.97	34%
2	604	604	32%	53%	.28	.10	.9	-1.36	.07	-.26	-.93	-1.36	-1.06	52%
3	623	623	33%	86%	.85	.99	1.2	.51	.06	1.49	.55	.51	.51	55%
4	258	258	14%	100%	1.99	1.97	1.0	2.35	.08	(3.54)	2.65	2.35	2.47	100%
										(Mean)		(Modal)	(Median)	

Figure 9.5

Probability Curb of the Rubric (Dataset 1: Levels 0–4)



The raters' results were similar across the datasets (see Table 9.6). For example, in Dataset 1 (see Figure 9.6), ChatGPT 1 constantly provided slightly more lenient scores than human raters (e.g., –0.58). Human raters varied slightly in rater severity estimates, ranging from –0.29–0.34), but within an acceptable range of –1 to +1. Infit MSs were between 0.75–1.17 among all the raters, with ChatGPT ideally close to 1 (0.91). Separation and strata were 3.28 and 4.71, meaning that raters were separated into three or four groups in severity differences; this was not ideal when raw ratings are used for high-stakes decision making, but acceptable when MFRM is used, which adjusts scores given rater severity differences. The exact agreement of all data included was higher than the agreement expected by MRFM (e.g., 63.5% > 38.1%; see Table 9.6). It is noteworthy that this study uses diverse raters, with and without deep content knowledge, to examine the applicability of the content rubric. The results show that the assessment of content in CLIL classrooms is feasible among diverse raters who may not be experts in the content field.

Figure 9.6

Rater Measurement Reports of Dataset 1 (Humans + ChatGPT 1)

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	N	Rater
553	260	2.13	2.26	.34	.08	.98	-.1	1.02	.2	.90	.50	59.1	37.7	5	RaterE
560	260	2.15	2.29	.29	.08	.97	-.3	.93	-.7	1.00	.44	61.4	37.9	4	RaterD
432	195	2.22	2.32	.24	.09	.75	-2.7	.77	-2.3	1.30	.50	68.1	38.4	2	RaterB
447	195	2.29	2.40	.11	.09	.92	-.8	.92	-.7	1.09	.50	65.0	38.8	6	RaterF
598	260	2.30	2.44	.04	.08	1.06	.6	1.02	.2	.93	.40	64.1	38.5	7	RaterG
628	260	2.42	2.55	-.16	.08	1.10	1.0	1.09	.9	.86	.47	64.7	38.6	1	RaterA
421	185	2.28	2.62	-.29	.10	1.17	1.5	1.12	1.1	.81	.40	72.4	38.1	3	RaterC
687	260	2.64	2.78	-.58	.09	.91	-1.0	.88	-1.3	1.16	.51	56.8	37.2	8	ChatGPT1
540.8	234.4	2.30	2.46	.00	.09	.98	-.2	.97	-.3		.46			Mean (Count: 8)	
92.0	33.2	.15	.17	.30	.01	.12	1.3	.11	1.1		.04			S.D. (Population)	
98.4	35.5	.16	.18	.32	.01	.13	1.4	.12	1.2		.05			S.D. (Sample)	

Model, Populn: RMSE .09 Adj (True) S.D. .29 Separation 3.28 Strata 4.71 Reliability (not inter-rater) .92
Model, Sample: RMSE .09 Adj (True) S.D. .31 Separation 3.53 Strata 5.04 Reliability (not inter-rater) .93
Model, Fixed (all same) chi-squared: 97.3 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-squared: 6.5 d.f.: 6 significance (probability): .37
Inter-Rater agreement opportunities: 5910 Exact agreements: 3753 = 63.5% Expected: 2252.9 = 38.1%

The bias analysis results in Table 9.7 revealed that bias interactions between students and tasks were the largest among all the bias interactions, followed by those between students and tasks and between students and criteria (e.g., 18.42%, 7.78%, and 5.73%, respectively in Dataset 1: Humans and ChatGPT 1). Table 9.8 showed that although there were biases in each analysis, the percentages of biases were similar across datasets (e.g., in the bias of students and tasks, 42.31–44.23%). Slight differences in the number of biases are mostly attributed to the larger degrees of freedom and greater power to reject a false null hypothesis owing to the addition of ChatGPT. Furthermore, rater-related biases were found among human raters, with the exception of biases between tasks and raters. In Datasets 1–3, each ChatGPT produced a lenient bias with Task 3, which was the easiest of the three tasks (see Appendixes C and D). One speculation for this is that Task 3 allowed students to generate a more structured and concrete detailed script because they were required to talk about one good learning material concretely in detail, rather than selecting one test or learning experience in Tasks 1 and 2. ChatGPT might have detected such task differences that led to script differences at the surface level, providing higher scores than human raters and causing bias.

Table 9.7

Percentages of Variances Explained by Rasch Measures and Bias

Dataset	0 (Humans)	1 (Humans+ ChatGPT1)	2 (Humans+ ChatGPT2)	3 (Humans+ ChatGPT3)
By Rasch measures	54.04	55.64	55.02	55.50
By Bias Interactions				
Student x task	18.90	18.42	19.35	19.07
Student x rater	7.70	7.78	7.74	7.80
Student x Criteria	6.39	5.75	5.60	5.42
Task x rater	1.29	1.31	1.26	1.47
Task x criteria	1.18	1.10	1.11	1.06
Rater x criteria	0.70	0.66	0.63	0.63

Table 9.8

Percentages of Biases

Dataset	0 (Humans)	1 (Humans+ ChatGPT1)	2 (Humans+ ChatGPT2)	3 (Humans+ ChatGPT3)
Student x task	42.31 (22/52)	44.23 (23/52)	44.23 (23/52)	42.31 (22/52)
Student x rater	1.68 (4/238)	1.82 (5/275)	1.45 (4/275)	1.45 (4/275)
Student x Criteria	3.78 (7/185)	4.86 (9/185)	4.86 (9/185)	4.32 (8/185)
Task x rater	14.29 (3/21)	20.83 (5/24)	16.67 (4/24)	25.00 (6/24)
Task x criteria	20.00 (3/15)	13.33 (2/15)	26.67 (4/15)	20.00 (3/15)
Rater x criteria	0.00 (0/35)	2.50 (1/40)	2.50 (1/40)	0.00 (0/40)

Discussion

The first RQ investigated the degree to which the ChatGPT ratings agreed with human ratings. Rater agreements, as measured by Fleiss' kappa, were not excellent, but acceptable across criteria and among human raters with and without ChatGPT ratings (see Table 9.5). Additionally, ChatGPT constantly produced higher interrater agreement than human raters (e.g., .71 vs. .46 in task achievement). MFRM results also showed acceptable overall agreement among raters, which was higher than the one expected by MFRM (see Table 9.6, Exact vs. expected agreement). The second RQ examined how ChatGPT ratings compared with human ratings in terms of severity and consistency. The results showed that the human and ChatGPT ratings were consistent, with minor variations in severity, with ChatGPT being slightly more lenient (see Table 9.6). Table 9.4 suggests that the human raters and ChatGPT had similar standard deviations, and that all the levels were used in the rubric. The third RQ investigated the degree to which ChatGPT exhibited biases toward test takers, tasks, and scoring criteria. Most biases were observed from human raters, with ChatGPT producing only one bias related to Task 3.

No bias between ChatGPT and the criteria indicates that ChatGPT was slightly lenient in all five criteria. Two reasons for this are explored. First, because ChatGPT is trained to avoid providing harmful responses (OpenAI, 2022) and scoring strictly can be considered harmful to

people, the AI's scores may have been lenient, despite a detailed prompt. Second, while ChatGPT uses an advanced large language model that enables the processing and generation of text, understanding the text context and integrating background information with text remains challenging (Bhattacharya et al., 2024), leading to difficulty in scoring content, particularly content accuracy. Figure 9.7 shows an example script that received more than one-level different scores between human raters and ChatGPT.

The current results are consistent with those of previous studies (e.g., Mizumoto & Eguchi, 2023), which showed moderate agreement between human and ChatGPT ratings. However, this study differs from Yamashita (2024), but aligns with Shin and Lee (2024) in that we found that ChatGPT (a) is slightly more lenient than human raters (see possible reasons above), (b) has similar score variations with human raters, and (c) has consistent scores that showed no overfit. While this study assesses content, deemed difficult to evaluate even by human raters by previous studies, we used a detailed prompt with sample scripts along with benchmarking scores, which enabled a more adequate assessment from a measurement perspective, consistent with Shin and Lee (2024).

Figure 9.7

Example Response for Task 2 (Modified Based on the Student's Original Script) on the Left and the Researchers' Interpretation of the ChatGPT's Evaluation of the Response on the Right

Let me now explain TOEIC. First, we have all taken different kinds of tests before. Recently, the TOEIC has become a very important test for getting a job. The TOEIC Speaking Test has five types of questions. The first is a reading aloud question. Next is a picture description question. After that, there is a response question, an informed response question, and finally, an opinion statement question. Each question has a time limit. Compared to the Eiken Speaking Test, the TOEIC is based on many different situations. Honestly, I think it is harder than the Eiken Speaking Test because *the questions are different from real-life situations.* However, it is important to improve your speaking skills. Out of the five question types, today I will talk about the photo description questions and the informed response questions. First, let's start with the Describing a Photo. In these questions, you look at a photo and describe what is happening. This tests your grammar and vocabulary. In this picture, a woman wearing blue is reading a book while talking. You can say, "The woman wearing blue is reading a book while talking." This may be hard, but important point is to use simple grammar and speak correct sentences. Second,	*Interpretation of differences between human and ChatGPT ratings* ChatGPT seems not to have captured aspects of the script that lower ratings. The script did not satisfy key elements to be included, which human raters noticed. *Task achievement*: It did not respond to the questions that should be included in the script (e.g., What test would you like to study to increase your speaking ability? Why do you want to study for the test?) *Logicality*: The script included discourse markers (e.g., *First*, *Honestly*, *However*), but the script ended half way, and the underlined sentences were illogical. The overall coherent structure is also missing in the script. *Content accuracy*: The script elucidated TOEIC Speaking test, but it did not include a scoring method or a method to conduct the test. Therefore, it lacked comprehensiveness of the content. Furthermore, the *italicized sentence* is not very accurate as the test covers everyday life as well as professional interactions. While human raters detected insufficiency in the script and assigned lower scores, ChatGPT failed to do so and provided higher scores.

Note. See Appendix E for more details.

Conclusion

This study examines ChatGPT's potential to yield scores comparable to ratings from human raters, which was generally supported, except for ChatGPT's slightly lenient and less biased scores. Accordingly, we consider the use of ChatGPT appropriate to complement human raters, and we posit the necessity of involving multiple humans or ChatGPT scoring, because the agreement was not strong.

This study has several limitations. First, we employed a detailed prompt that yielded results consistent with Shin and Lee's (2024). However, we did not undertake a direct comparison between a detailed prompt and a zero-shot prompt. Such a comparison could reveal the impact of prompt detailedness. Second, we analyzed relatively small datasets. Additionally, although we provided three to five benchmark scripts and their human-rated scores for each task, more sample scripts, scores, and scoring trials of ChatGPT may yield more precise and representative results. We also did not include non-Japanese raters and including them may present a more comprehensive picture of comparing human and ChatGPT scores. Furthermore, AI tools other than ChatGPT should be tested to explore possibilities of using AI to score content. Third, this study focused on quantitative methods; using a mixed-methods approach using qualitative elicitation and analytical methods of rater behavior and perception would reveal a general trend of and processes behind scoring content in CLIL classrooms.

Despite these limitations, the scores were generally relevant to understanding the degree to which ChatGPT can be used with human raters in low-stakes situations (Xi, 2023). This study demonstrates that assessing content in CLIL classrooms can be managed by creating a rubric and other detailed supplementary scoring information, along with training for human raters and detailed prompts for ChatGPT. This study shows that Fleiss' kappa and MFRM allow for more detailed and transparent results for nonexperts and supports the usefulness of MFRM to evaluate human and AI-based automated scoring (Shin & Lee, 2024; Yamashita, 2024).

Acknowledgments

The current study was funded by JSPS Grants-in-Aid for Scientific Research 24K00082, 24K16120, 24K22785, and 24K00081. We would like to appreciate those who helped us score scripts and improve our manuscript, particularly, Yumi Koyamada, Hiroto Takamure, Runo Inagawa, Haruno Koshida, and the audience in the presentation. Supplementary Materials can be found at https://osf.io/w7shf/

References

Bhattacharya, P., Prasad, V. K., Verma, A., Gupta, D., Sapsomboon, A., Viriyasitavat, W., & Dhiman, G. (2024). Demystifying ChatGPT: An in-depth survey of OpenAI's robust large language models. *Archives of Computational Methods in Engineering*, *31*, 4557–4600. https://doi.org/10.1007/s11831-024-10115-5

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

deBoer, M., & Leontjev, D. (Eds.). (2020). *Assessment and learning in Content and Language Integrated Learning (CLIL) classrooms*. Springer. https://doi.org/10.1007/978-3-030-54128-6

DeMars, C. E. (2021). Violation of conditional independence in the many-facets Rasch model. *Applied Measurement in Education*, *34*(2), 122–138. https://doi.org/10.1080/08957347.2021.1890743

Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 73–95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.

Linacre, J. M. (2024a). Facets: Many-facet Rasch measurement (Version 4.2.3) [Computer software]. MESA Press. https://www.winsteps.com/facets.htm

Linacre, J. M. (2024b). *A user's guide to FACETS (64-bit) Rasch-model computer programs: Program manual 4.2.3*. https://www.winsteps.com/manuals.htm

McNamara, T., Knoch, T., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford University Press.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

OpenAI. (2022, November 30). *Introducing ChatGPT*. Retrieved January 12, 2025, from https://openai.com/index/chatgpt/

Sato, T. (2024). Assessing the content quality of essays in content and language integrated learning: Exploring the construct from subject specialists' perspectives. *Language Testing*, *41*(2), 316–337. https://doi.org/10.1177/02655322231190058

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies, 29*, 24735–24757. https://doi.org/10.1007/s10639-024-12817-6

Xi, X. (2023). Advancing language assessment with AI and ML—Leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, *20*(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488

Xu, Y., Polio, C., & Pfau, A. (2024). Optimizing AI for assessing L2 writing accuracy: An exploration of temperatures and prompts. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 151–174). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.10

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics*, *3*(3), 100133. https://doi.org/10.1016/j.rmal.2024.100133

Appendix A

A Prompt for Task 1 for ChatGPT (Benchmarking essays were paraphrased while maintaining errors and structures)

I would like you to mark a presentation script written by Japanese learners of English as a foreign language (EFL) learners. Each script is scored on a scale of 0 to 4, with 4 being the highest and 0 the lowest. You need to give me the reason why you assign that specific score. Report both the scores and the reasons.

The directions for the students are as follows:

Write a presentation script. You have 2 min (between 1 min and 3 min) to make a presentation.

Here are example questions that you can answer in the talk.
a) What is the effective experience or activity that increased your English ability?
b) What is the ineffective experience or activity that did not increase your English ability?
c) Why were they effective or ineffective?
d) In the future, what activities are you going to do to increase your English ability?

Points to include: (Note: Students should include both Topics 1 and 2)
Introduction: Say your name and what you mainly talk about. (Don't deduct points even when students did not say their names)

Topic 1: Talk about your English (or another foreign language) learning experiences. Were they effective? Why or why not?
Use the information below: Input / intake / output / pushed output / motivation / anxiety (Note: Students learned these terms related to theories of second language acquisition.)
•Experience 1:
•Experience 2:

Topic 2: How can you improve your L2 (second language) learning?
Use the information below: Input / intake / output / pushed output / motivation / anxiety (Note: Students learned these terms related to theories of second language acquisition.)
•Method 1:
•Method 2:

Conclusion: Summarize your talk briefly.

The script is scored based on the following rubric:

Task Achievement: The extent to which the script answers the question and is relevant to the topic.
[score 4] - Provides a sufficient response to the writing prompt and writes extensively on the topic.
[score 3] - Provides a mostly sufficient response to the prompt and writes largely on the topic.
[score 2] - Provides a partial response to the prompt and writes partially on the topic.

[score 1] - Does not respond to the prompt and does not write on the topic.
[score 0] - Too short to assess.

Clarity of Content: The extent to which the content, message, and meaning are clear and comprehensible. Scores are deducted if there are errors that hinder understanding.
Errors related to vocabulary and grammar are closely related to clarity. If the errors do not hinder understanding (i.e., minor errors), there is no need to lower the score. Even if there are many minor errors, as long as each one does not hinder understanding, it is not considered problematic.
Excluded elements: Errors in punctuation. Since the document was originally created as a script for presentation, the punctuation has been revised throughout. Additionally, unnecessary spaces may be present due to issues related to the Learning Management System (LMS), which may not be the fault of the individual examinees.
[score 4] The content (including message and meaning) is sufficiently clear.
[score 3] The content is moderately clear.
[score 2] The content is partially clear.
[score 1] The content is unclear.
[score 0] Too short to assess.

Elaboration: The extent to which the explanations, examples, and reasons are present and sufficient.
[score 4] There are explanations, examples, and reasons, and they are sufficiently detailed (with sufficient quantity).
[score 3] There are explanations, examples, and reasons, and the level of detail is moderately high.
[score 2] There are explanations, examples, and reasons, but they are only partially detailed.
[score 1] There are no explanations, examples, or reasons.
[score 0] Too short to assess.

Logicality: The extent to which the flow of ideas is logical, the relationships between paragraphs are strong, and the quality of reasoning is high.
Logicality: Consistency in the flow of ideas and paragraphs. Appropriateness of the location where ideas and paragraphs are written, and the overall structure (use of discourse markers such as "First," "In conclusion," etc.). Presentation of the overall flow at the beginning. Connection between opinions and supporting reasons or details. Convincing reasoning behind opinions. Absence of sudden topic changes, content fluctuations, or unnecessary repetitions.
[score 4] The flow of ideas or paragraphs is sufficiently logical, and paragraphs are sufficiently linked.
[score 3] The flow of ideas or paragraphs is moderately logical, and paragraphs are fairly linked.
[score 2] The flow of ideas or paragraphs is only partially logical, and paragraphs are only partially linked.
[score 1] The flow of ideas is not logical, and paragraphs are not linked.
[score 0] Too short to assess.

Accuracy of Content: The degree to which the ideas, information, premises are correct and comprehensive.

It is important to ensure that the content aligns with theories of second language acquisition and language testing research, as well as with general facts. Additionally, during the 2-minute presentation, it is essential to cover the topic comprehensively.
Comprehensiveness includes incorporating elements of "depth and complexity of thought," which involves considering and including ideas from multiple perspectives.
[score 4] The ideas, information, and premises are sufficiently accurate and comprehensive.
[score 3] The ideas, information, and premises are moderately accurate and comprehensive.
[score 2] The ideas, information, and premises are only partially accurate and comprehensive.
[score 1] The ideas, information, and premises are not accurate or comprehensive.
[score 0] Too short to assess.

Note: 4 = A; 3 = B+; 2 = B; 1 = C (in line with the national research reference materials). Generally, a score of 2 = B is considered a passing level.

Difference between 3 and 2: A score of 3 indicates that there are some (not many) areas where the criteria are not met, while a score of 2 indicates that there are many areas where the criteria are not met.

Below is the data for you to familiarize yourself with scoring using the topic and the rubric above:

Training 1: Data 1: No. X *(Data 1 = Task 1)*

I speak information my learning English and how I can improve. It's input and output.

This response is scored 1, 1, 0, 0, and 0 for Task Achievement, Clarity of Content, Elaboration, Logicality, and Accuracy of Content, respectively.

Training 2: Data 1: No. Y

Hello everyone. I'm * *. Today, I will discuss my experiences of learning English. First, I will … *(omitted in this article; Not omitted for ChatGPT)*

This response is scored 3, 3, 2, 3, and 2 for Task Achievement, Clarity of Content, Elaboration, Logicality, and Accuracy of Content, respectively.

(Another essay)

This response is scored 3, 3, 2, 3, and 2 for Task Achievement, Clarity of Content, Elaboration, Logicality, and Accuracy of Content, respectively.

Now, let's start to score the following script. There are five criteria: Task Achievement, Clarity of Content, Elaboration, Logicality, and Accuracy of Content. Please give a score for each criterion and give me the reason why you assign that specific score. Report both the scores and the reasons. Please save the scores and reasons for each script in Microsoft Excel format and export it. Here is the template for the output.

	Score					Reason				
Script No.	Task Achievement	Clarity of Content	Elaboration	Logicality	Accuracy of Content	Task Achievement	Clarity of Content	Elaboration	Logicality	Accuracy of Content

Appendix B

Figure B1

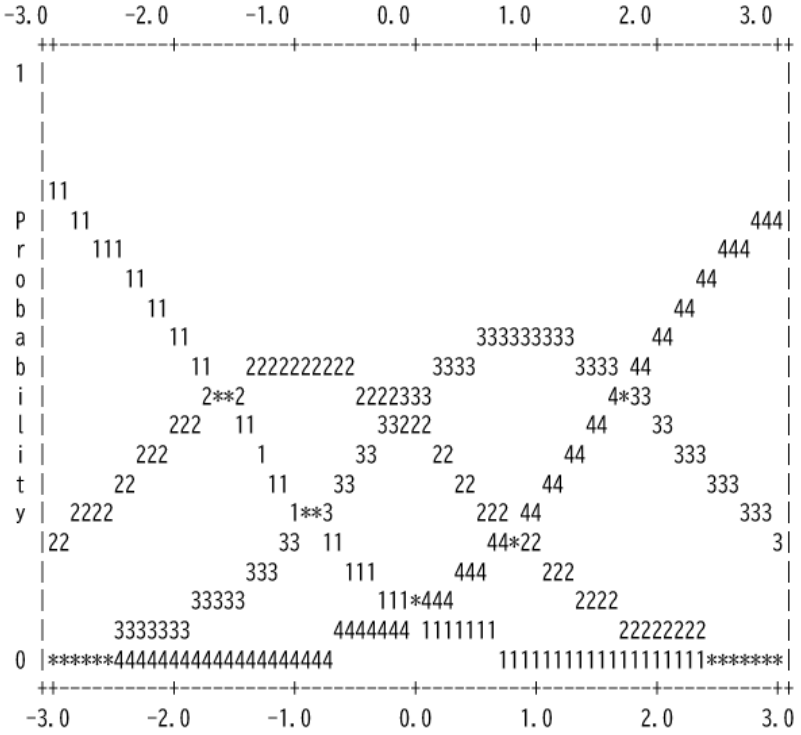
Category Statistics of the Revised Rubric (Levels 1–4, with Levels 0 and 1 combined; Dataset 1: Humans + ChatGPT 1)

DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	390	320	18%	18%	-1.53	-1.40	.9			(-2.82)		low	low	100%
2	604	604	33%	51%	-.38	-.54	1.0	-1.61	.07	-.92	-1.98	-1.61	-1.79	50%
3	623	623	35%	86%	.26	.37	1.1	-.12	.06	.89	-.04	-.12	-.07	55%
4	258	258	14%	100%	1.40	1.34	1.0	1.73	.08	(2.91)	2.03	1.73	1.85	100%
										(Mean)		(Modal)	(Median)	

Note. Total and Used category counts are different in number (i.e., 390 and 320) because students with all minimum scores (Level 1) were excluded in the analysis.

Figure B2

Probability Curb of the Revised Rubric (Dataset 1: Levels 1–4 on the right)



Appendix C

Table C1

Biases Found between Tasks and Raters across Datasets

Dataset 0 (Human)	Dataset 1 (Humans+ ChatGPT 1)	Dataset 2 (Humans+ ChatGPT 2)	Dataset 3 (Humans+ ChatGPT 3)
T3 x Rater E (L) T2 x Rater E (S) T3 x Rater G (S)	+ T3 x ChatGPT 1 (L) + T2 x Rater D (L)	+ T3 x ChatGPT 2 (L)	+ T3 x ChatGPT 3 (L) + T2 x Rater D (L) + T3 x Rater B (S)

Note. T = Task. L = Lenient. S = Strict.

Appendix D

Table D1

Pairwise Report of Bias Interactions between Tasks and Raters

Target No.	Target-Rater	Measre	S.E.	Obs-Exp Average	Context No.	Target-Task	Measre	S.E.	Obs-Exp Average	Context No.	Task	Target-Contrast	Joint S.E.	Rasch-Welch *t*	*df*	*p*
Humans only																
5	RaterE	0.62	0.13	-0.23	2	T2	-0.27	0.14	0.28	3	T3	0.89	0.20	4.56	182	< .001
5	RaterE	0.38	0.14	-0.09	1	T1	-0.27	0.14	0.28	3	T3	0.66	0.20	3.24	165	.002
7	RaterG	-0.23	0.14	0.12	1	T1	0.24	0.14	-0.15	3	T3	-0.46	0.20	-2.34	163	.020
2	RaterB	-0.04	0.14	0.12	1	T1	0.35	0.14	-0.12	3	T3	-0.39	0.20	-2.00	163	.048
Humans + ChatGPT 1																
5	RaterE	0.65	0.13	-0.2	2	T2	-0.14	0.15	0.25	3	T3	0.80	0.20	4.04	182	< .001
5	RaterE	0.46	0.14	-0.08	1	T1	-0.14	0.15	0.25	3	T3	0.60	0.20	2.96	166	.004
8	GPT1	-0.31	0.14	-0.14	2	T2	-0.94	0.16	0.15	3	T3	0.63	0.22	2.91	181	.004
7	RaterG	-0.15	0.14	0.13	1	T1	0.38	0.14	-0.18	3	T3	-0.53	0.20	-2.66	164	.009
4	RaterD	-0.01	0.14	0.18	2	T2	0.47	0.14	-0.10	3	T3	-0.48	0.20	-2.48	182	.014
7	RaterG	-0.11	0.14	0.09	2	T2	0.38	0.14	-0.18	3	T3	-0.49	0.20	-2.48	182	.014
2	RaterB	0.03	0.14	0.14	1	T1	0.49	0.14	-0.14	3	T3	-0.46	0.20	-2.32	163	.022
4	RaterD	0.42	0.14	-0.08	1	T1	-0.01	0.14	0.18	2	T2	0.43	0.20	2.16	160	.032

Target No.	Target-Rater	Measre	S.E.	Obs-Exp Average	Context No.	Target-Task	Measre	S.E.	Obs-Exp Average	Context No.	Task	Target-Contrast	Joint S.E.	Rasch-Welch *t*	*df*	*P*
Humans + ChatGPT 2																
5	RaterE	0.66	0.13	-0.21	2	T2	-0.16	0.14	0.25	3	T3	0.82	0.20	4.20	182	< .001
5	RaterE	0.42	0.14	-0.07	1	T1	-0.16	0.14	0.25	3	T3	0.58	0.20	2.87	166	.005
7	RaterG	-0.18	0.14	0.14	1	T1	0.35	0.14	-0.18	3	T3	-0.53	0.20	-2.70	164	.008
2	RaterB	0.00	0.14	0.14	1	T1	0.47	0.14	-0.14	3	T3	-0.46	0.20	-2.37	163	.019
9	GPT2	-0.30	0.14	-0.11	1	T1	-0.80	0.16	0.15	3	T3	0.50	0.21	2.35	167	.020
4	RaterD	0.01	0.14	0.16	2	T2	0.45	0.14	-0.10	3	T3	-0.44	0.19	-2.27	182	.024
7	RaterG	-0.09	0.14	0.07	2	T2	0.35	0.14	-0.18	3	T3	-0.44	0.20	-2.27	182	.024
9	GPT2	-0.35	0.14	-0.06	2	T2	-0.80	0.16	0.15	3	T3	0.45	0.21	2.11	182	.036
Humans + ChatGPT 3																
5	RaterE	0.66	0.13	-0.20	2	T2	-0.12	0.15	0.24	3	T3	0.78	0.20	3.95	182	< .001
10	GPT3	-0.35	0.14	-0.15	1	T1	-1.15	0.17	0.23	3	T3	0.80	0.22	3.62	167	< .001
10	GPT3	-0.37	0.14	-0.12	2	T2	-1.15	0.17	0.23	3	T3	0.78	0.22	3.53	179	.001
7	RaterG	-0.17	0.14	0.14	1	T1	0.40	0.14	-0.19	3	T3	-0.57	0.20	-2.90	164	.004
5	RaterE	0.43	0.14	-0.06	1	T1	-0.12	0.15	0.24	3	T3	0.55	0.20	2.69	166	.008
2	RaterB	0.01	0.14	0.16	1	T1	0.51	0.14	-0.15	3	T3	-0.51	0.20	-2.57	163	.011
4	RaterD	0.00	0.14	0.17	2	T2	0.50	0.14	-0.11	3	T3	-0.49	0.19	-2.54	182	.012
7	RaterG	-0.09	0.14	0.08	2	T2	0.40	0.14	-0.19	3	T3	-0.50	0.20	-2.54	182	.012

Note. GPT = ChatGPT.

Appendix E

Supplementary Information for Figure 9.7 (Example Response for Group 1, Task 2)

Results from Dataset 0 (Humans)

Ability estimates = 0.65 logits, Standard error = 0.16, Infit mean squares = 1.06, Outfit mean squares = 1.13

Table E1

Mean Ratings by Human Raters and ChatGPT-4

	Task Achievement	Content clarity	Elaboration	Logicality	Content accuracy
Human ratings	1.60	2.60	2.00	1.60	2.20
ChatGPT-4 rating	3.00	3.00	2.33	3.00	3.33
Difference	–1.40	–0.40	–0.33	–1.40	–1.13

Human raters' comments (translated by the first author)

Task achievement and *logicality*: Rater E "Redundant, with little connection to personal relevance."
Elaboration: Rater C "I can somewhat imagine what kind of test it is, but I don't know the details."
Logicality: Rater D "It was hard to choose between Levels 1 or 2. … the flow was not indicated. Toward the latter part, it felt like the explanation of the test was merely listed without coherence."
Elaboration and *logicality*: Rater G "The latter part becomes a simple listing of each test part's content, which I think lacks detail."

Comments from ChatGPT 2

Task achievement: Provides a mostly sufficient response to the prompt, explaining the choice of TOEIC and detailing its format.
Clarity of content: Content is generally clear and comprehensible, with minor issues that do not significantly hinder understanding.
Elaboration: Provides some explanations and reasons but lacks depth in detailing the test's sections and personal motivation.
Logicality: The flow of ideas is moderately logical, with a clear structure and reasonable progression from introduction to details.
Content accuracy: Information is generally accurate, focusing on TOEIC's format and use, but lacks comprehensive details on speaking ability measurement.

About the Authors

Rie Koizumi is a Professor of Applied Linguistics at the University of Tsukuba, Japan. Her research focuses on test validation, speaking assessment, and skill profiles. An example article was published in *Validity argument in language testing: Case studies of argument-based validation research* (Chapelle & Voss [Eds.], 2022, Cambridge University Press), *Language Assessment Quarterly*, *System*, *Language Testing in Asia*, and other journals. She co-hosted the online Language Testing Research Colloquium (Tokyo, 2022). She has received the Japan Language Testing Association (JLTA) Best Book Award (2021, 2022, and 2024) and the JLTA Best Paper Award (2023). She served as a board member of JLTA since 2008 and as Secretary General from 2014 to 2020. She is on the editorial board of *Language Assessment Quarterly* (2023–present). For more information, visit her website: https://trios.tsukuba.ac.jp/en/researcher/0000004741

Saki Suemori is an Assistant Professor of Applied Linguistics at the University of Tsukuba, Japan. Her research interests include teacher psychology and CLIL, and has published papers in *JALT Journal*. She is particularly interested in how teacher psychology influences learner psychology. She has been conducting research primarily using a qualitative approach and has co-edited a book on qualitative research methodologies. She is also interested in how qualitative research has been applied in the field of language teaching and learning. Since 2016, she has served on the editorial committee of Japan CLIL Pedagogy Association (J-CLIL), and is currently developing a CLIL textbook for Japanese medical students in Japan. For more information, visit the following website: https://trios.tsukuba.ac.jp/en/researcher/0000004599

Yusuke Kubo is a PhD student in the English Language Education Program at the University of Tsukuba and an Associate Professor in the Faculty of Education at the University of Yamanashi, Japan. His research interests include language testing and assessment, with a focus on evaluating content-related aspects and thinking skills through productive and receptive tasks. His current research explores the assessment and diagnosis of critical thinking skills using a multiple-choice reading test using an argument-based validation framework. Additionally, he investigates how functional adequacy impacts the quality of writing products, focusing on the use of cohesive devices. For contributing to this field, he has received the Japan Association for Language Education and Technology Kanto Chapter Encouragement Award (2024).

Chapter 10

A Validation Study of AI-Generated Prompts in CILS (Certification of Italian as a Foreign Language) B2 Exams

Giulia Peri, Sabrina Machetti, and Paola Masillo
University for Foreigners of Siena

This study presents an ongoing validation analysis of generative AI (ChatGPT-4; OpenAI 2023) in creating test prompts for the written production component of the CILS (Certification of Italian as a Foreign Language) B2 exam. As AI technologies increasingly influence language assessment and test development, a key challenge is ensuring that AI-generated writing test prompts maintain validity, appropriateness, and alignment with assessment constructs. This issue is particularly relevant for high-stakes language certifications, where the demand for new test items is growing, but the number of trained item writers remains limited. The research aims to evaluate ChatGPT-4's capability to generate writing test prompts reflecting the CILS B2 target domain and to investigate how these prompts are perceived by CILS experts and test-takers. Following an argument-based validity framework (Chapelle & Voss, 2021), the study employs a mixed-methods approach to collect evidence for the domain definition inference.

Introduction

The Certification of Italian as a Foreign Language (*Certificazione di Italiano come Lingua Straniera* - CILS) is an Italian language qualification offered by the CILS Center of the University for Foreigners of Siena (Italy). CILS exams are standardized tests of Italian measuring general language proficiency, are aligned with the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* ('CEFR'; CoE, 2001) and are staged

Cite as: Peri, G., Machetti, S., & Masillo, P. (2025). A validation study of AI-generated prompts in CILS (Certification of Italian as a Foreign Language) B2 exams. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 197-218). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.10
©2025 Peri, Machetti, and Masillo. Published under a CC BY license.

on the six CEFR levels (A1 to C2). The target audience for the CILS exams consists of a general adult population with an average cultural background who study the Italian language for educational, professional, or general cultural purposes, both in Italy and abroad.

CILS exams are recognized by universities, employers, and institutions worldwide. Since their introduction in 1993 and over 30 years of activity, approximately 400,000 test-takers have taken the CILS exams in Italy and 87 foreign countries. Currently, all CILS exams are in paper format. Therefore, they are administered in person at CILS Test Centers and are entirely assessed centrally at the CILS Center in Siena (Machetti, 2022). As stated in the CILS Guidelines (Centro CILS, 2009), CILS exams are text and task based. Thus, they present test takers with tasks that relate to situations representative of real-life ones, with the aim of measuring the ability of test-takers to use the Italian language in various domains and contexts. Each CILS exam consists of five components, measuring the following linguistic and communicative skills: listening comprehension, reading comprehension, use of Italian, written and oral production and interaction. Proficiency in Italian written and oral production and interaction is tested through extended-production tasks, which prompt linguistic performance in authentic contexts (CoE, 2001). Task difficulty varies based on the CEFR indications, with moderating factors such as time constraints and type of support provided (Centro CILS, 2009). Regarding the scoring of extended-production tasks in the written component of CILS exams, trained and standardized raters assess test-takers performances based on rating criteria and elaborated rating scales with descriptors, developed according to the proficiency framework. These criteria include communicative effectiveness, register and stylistic appropriateness, morphosyntactic accuracy, lexical adequacy and richness, as well as spelling and punctuation.

The entire CILS test cycle occurs with the support of basic information technology at every stage (e.g., computers, internet, WordOffice, SPSS). However, as the rapid expansion of technology-mediated language assessment develops, test developers are facing the challenge of producing high-quality items efficiently and cost-effectively, stressing the need for practical solutions without impacting negatively on validity (Rossi, 2023). More specifically, the CILS Centre faces challenges related to the increasing workload and the scarcity of well-trained item writers (Gallina, Machetti, & Masillo, 2019). Moreover, the rapid development of artificial intelligence (AI), its spread in the language testing field (Voss, 2024) and changing learner needs prompts the alignment of the assessment process with contemporary technological and educational needs.

Among the most notable advancements in AI-driven language tools is ChatGPT, an advanced model developed by OpenAI (OpenAI, 2023). Built on the Generative Pre-trained Transformer (GPT) architecture, it excels in conversation, content creation, coding, and tutoring. Pre-trained on vast text data and fine-tuned with human feedback, it generates coherent and context-aware responses during interactions. In Italy, the use of ChatGPT began to spread primarily from 2023 onward, with usage mainly linked to informal contexts. A particularly significant aspect of this situation is that ChatGPT was also at the center of much controversy, especially in 2023 and particularly in Italy, where it was blocked by the Italian Data Protection Authority due to privacy concerns. The blocking of ChatGPT may have contributed, directly or

indirectly, to the development of a more skeptical or negative perception of AI tools among both professionals and the public. On the other hand, the challenge of developing and fine-tuning AI technologies for non-English languages is a critical issue in the global deployment of AI solutions (Robinson et al., 2023).

The predominance of English in AI development not only impacts the effectiveness of the tool in other linguistic contexts but also its cultural sensitivity and relevance. For instance, English-speaking countries have explored the topic of AI for education more extensively, with cutting-edge studies and practical applications leading the way. For instance, many are the studies looking at the interrelationships between language assessment and AI (e.g., Van Moere & Downey, 2016; Voss et al., 2023; Xi, 2023) or item development and AI (e.g., Attali et al., 2022; Gierl et al., 2012). In the Italian context, research on AI-driven language learning, teaching and assessment is still emerging (e.g., in the field of learning and teaching, see Cinganotto, Sbardella & Montanucci, 2024). However, even in contexts outside of Italy, there appears to be a lack of published studies specifically focusing on the use of AI to generate writing test prompts for written production tasks in the context of formal testing. In this context, the main challenge for the CILS Center is maintaining a high level of validity in its tests while ensuring greater practicality and operational sustainability, qualities that contribute to the usefulness of the test (Bachman & Palmer, 1996). This calls for a rethinking of the test design and administration methods, ensuring that advancements in AI and technology are leveraged effectively without compromising assessment quality.

In response to these challenges, our study explores the feasibility of using ChatGPT-4 to generate writing test prompts for the CILS exams. To ensure a rigorous validation process, this study adopts an argument-based validity framework (Chapelle, Enright, & Jamieson, 2008; Chapelle & Voss, 2021), focusing on the domain definition inference, as previously investigated by research in the English context (e.g., Jun, 2021; Knoch & Chapelle, 2018; Lee, 2023).

The argument-based validity framework is based on the "construction of an interpretative argument that lays out the grounds, inferences, warrants, and claims, all of which provide the foundation for proposed score interpretation and use" (Chapelle, Enright, & Jamieson, 2008, p. 23). At the foundation of the argument is the domain definition inference that establishes the link "between performances in the target domain to the observations of performance in the test domain" (Chapelle, 2012, p. 22; Chapelle, Enright, & Jamieson, 2008, p. 14). As for the "conceptual tools for developing validity arguments" (Chapelle, 2012, p. 20), the warrant supporting the domain definition inference is that observations of performance on the test are representative of performances in the target domain of language use (Chapelle, Enright, & Jamieson, 2008, p. 17). In turn, the warrant relies on two assumptions: (a) that assessment task representative of the target domain can be identified; and (b) that such assessment tasks can be simulated as test tasks. The backing for these assumptions involves a thorough domain analysis and meticulous task construction, as well as evaluations of the success of these processes' outcomes (Chapelle, Enright, & Jamieson, 2008, pp. 17-21). In the writing assessment, investigation of the domain definition inference reveals whether the writing test prompts elicit test-takers' performances that are representative of their writing performances in the target

domain. This is a fundamental concern for all test prompts, as their validity depends on whether they effectively assess the intended construct. However, when writing test prompts are generated by AI rather than human item writers, new factors must be considered, and they must be evaluated to determine whether they align with the established assessment construct and proficiency level.

Given these considerations, our study seeks to collect empirical evidence for the domain definition inference, examining whether AI-generated writing test prompts are representative of the CILS B2 writing test target domain and can be used as writing test prompts for the tasks in the CILS B2 written exam. The domain definition inference is based on the warrant that observations of performance on AI-generated writing test prompts in the CILS B2 exam are representative of the writing skills required at the CEFR B2 level, to which this CILS exam is aligned. This warrant is grounded in three key assumptions:

1. The CILS B2 construct for written interaction and production is well-defined.
2. AI-generated writing test prompts can be designed to be representative of the CILS B2 writing skills (RQ1).
3. AI-generated writing test prompts are appropriate for eliciting B2-level written interaction and production (RQ2).

To support the first assumption, a detailed analysis was conducted to examine the domain relevant to the written interaction and production required for the CILS B2 exam, drawing on the CILS Guidelines (2009) and CEFR B2 descriptors (CoE, 2001). This process confirmed the already established construct definition for written interaction and production tasks at the CILS B2 level. An overview of the construct definition is provided in Appendix A. Since the construct was already explicitly outlined, it did not require further empirical validation in this study.

For the second assumption, empirical validation was necessary to determine whether AI-generated prompts can be identified as representative of the target domain, the CILS B2 writing skills. To assess this, experts were involved in the evaluation of AI-generated writing test prompts, serving as backing for (RQ1). For the third assumption, the study examined test-taker perceptions of AI-generated prompts in terms of difficulty, clarity, and relevance (RQ2).

Therefore, to achieve the goal of this study, the following research questions were formulated stemming from the assumptions illustrated above:

1. To what extent does ChatGPT-4 generate writing test prompts that are perceived as representative of the CILS B2 target domain, according to expert evaluation?
2. How do test-takers perceive ChatGPT-4-generated writing test prompts compared to those created by human item writers in terms of difficulty, clarity, and relevance to the CILS B2 target domain?

Methods

Research Context and Design

This study was conducted using a mixed-method approach (see Instruments for more details) within the context of the CILS B2, which assesses test-takers 'proficiency in Italian as a S/FL at the CEFR B2 level. The exam tests multiple skills (see Introduction), including written interaction and production, organized in two tasks (see Appendix A). As already mentioned in the Introduction, the study used an argument-based approach referencing Chapelle and Voss (2021), with a specific focus on collecting evidence for the domain definition inference (Chapelle & Voss 2021, p. 35).

Participants

CILS Item Writers

The study involved seven Subject-Matter Experts (SMEs) specializing in Second/Foreign Language (S/FL) Italian Language Testing and Assessment. These experts were experienced CILS item writers who had been working for the CILS Center in various senior roles for 7 to 30 years. Their academic qualifications ranged from Master's degrees to PhDs, and they had extensive experience in language test development, task design, and proficiency assessment, particularly within the CILS certification framework.

Test-Takers

The participant sample consisted of 63 test-takers, aged 16 to18, enrolled in an Italian high-school in Istanbul (Türkiye). A purposive sampling method was adopted, as test-takers were specifically selected based on their upcoming participation in the CILS B2 exam in one of the forthcoming sessions organized by the CILS Test Center in Istanbul. This sampling strategy ensured that participants had a relevant proficiency level (CEFR B1 or B2) and a real test-taking motivation, making them suitable for evaluating the AI-generated and human-created writing test prompts.

Instruments

The study employed a set of instruments designed to address the research questions (RQs) by collecting backing (empirical evidence) to support or challenge the assumptions underlying the warrant that AI-generated writing test prompts are representative of the target domain.

To address RQ1, we employed the following instruments:

- Two writing prompt models for ChatGPT-4: Predefined writing task prompt models from a previous CILS B2 exam administration (one for Task 1 and one for Task 2) were used as reference for writing test prompt generation.
- AI-generated writing test prompts: Writing test prompts (no. 20) for both Task 1 and Task 2 were generated using ChatGPT-4, following structured input parameters describing CILS B2 human-made writing tasks.
- Expert survey (Survey 1): Survey 1 was designed to gather SMEs' evaluations of AI-generated writing test prompts, focusing on their appropriateness, alignment with the B2

construct, and AI detectability. The survey was administered online and structured into two sections, corresponding to the two writing tasks (Task 1 and Task 2) of the CILS B2 exam. The survey included multiple-selection questions for evaluating the writing test prompts (1-10) and open-ended responses for justifying each selection or give additional comments. Experts were unaware that all prompts had been generated by AI.

To address RQ2, we used the following instruments:

- Human-generated and AI-generated tasks: both human-created writing test prompts (i.e., the writing test prompt models) and two expert-selected AI-generated writing test prompts (one for Task 1 and one for Task 2) were administered to test-takers in a paper-and-pencil format under standardized test conditions. A total of four writing test prompts were administered (two for Task 1 and two for Task 2).

- Test-taker survey (Survey 2): Survey 2 was designed to gather test-taker perceptions after administration regarding the difficulty, clarity and relevance of writing test prompts without revealing whether they were AI- or human-generated. The survey was administered in-person and structured into two sections, corresponding to the two writing tasks (Task 1 and Task 2) in the CILS B2 exam. Each section contained multiple-choice questions (i.e., Likert scale, Yes/No) and open-ended responses, allowing participants to provide both quantitative ratings and qualitative insights on the prompts they reviewed.

Data Collection and Analysis

To collect backing for the domain definition inference, this study adopted a mixed-methods approach, combining descriptive frequency analysis of survey responses with content analysis of participants' evaluations. Data collection was structured to align with the two research questions, ensuring that both SMEs assessments (RQ1) and test-taker perceptions (RQ2) were systematically examined.

RQ1. To What Extent does ChatGPT-4 Generate Writing Test Prompts that are Perceived as Representative for the CILS B2 Domain, according to Expert Evaluation?

To answer this first research question, it was necessary to produce the main materials for the study, namely the writing test prompts. Before generating the writing test prompts with AI, a writing test prompt model from tasks developed by humans for the AI to use as an example was identified (Attali et al., 2022; Bejar, 2002). Two specific writing test prompts, named "Task 1" and "Task 2", were selected from a previous session of the CILS B2 exam (see Appendix B). This choice was guided by post-test statistics, routinely calculated by the CILS Center, which indicated the high quality of these writing test prompts in terms of validity and alignment with the measured construct. The study proceeded with the AI generation of 10 writing test prompts for Task 1 and 10 writing test prompts for Task 2, combining a model-based approach and a rule-based approach methodology (adapted from Kurdi et al., 2020) with ChatGPT-4. The rules provided to the AI through prompts referred to the CEFR B2 descriptors and to the structures identified by the CILS Guidelines for B2-level written interaction and production. Temperature settings were not modified, meaning ChatGPT-4 operated with its default deterministic

configuration to maintain output consistency and reproducibility across multiple prompt generations. Since ChatGPT is known to reflect biases present in its training data (Liu et al., 2024), precautions were taken to prevent the inclusion of inappropriate or sensitive content in the AI-generated writing test prompts. Specifically, the AI was instructed to adhere to the "CILS taboo list" (internal document), which excludes topics related to crimes, war, physical and mental health, etc.

A survey for SMEs (Survey 1) was developed to elicit the SMEs assessment of the AI-generated writing test prompts' adequacy and their detection of any perceived artificial intelligence involvement. Specifically, the survey aimed to gather SMEs input on the appropriateness and authenticity of writing test prompts for assessing writing ability at the B2 level of the CILS exam. SMEs first selected which writing test prompts they considered suitable for Task 1, evaluating aspects such as thematic relevance, difficulty, and alignment with CILS standards. They were asked to explain their selections, providing qualitative insights into the criteria they used. SMEs were then asked to identify which writing test prompts they believed were generated with the support of AI, without knowing that all writing test prompts were AI-generated. They were also invited to explain their reasoning, offering valuable qualitative data on the features or characteristics they associated with AI involvement. These questions were consistently applied to both Task 1 and Task 2 to allow for a comprehensive evaluation. Therefore, SMEs provided quantitative and qualitative feedback to investigate whether they saw AI-generated writing test prompts aligning with the B2 construct (Assumption 2).

RQ2. How do Test-takers Perceive ChatGPT-4-generated Writing Test Prompts Compared to Those Created by Human Item Writers in terms of Difficulty, Clarity, and Relevance to the CILS B2 Target Domain?

Two writing test prompts (one for Task 1 and one for Task 2) were selected from AI-generated outputs by two SMEs (see Appendix B), who conducted a preliminary review to ensure that the writing test prompts did not exhibit substantial divergence from CILS B2 requirements or present issues impeding test takers' performance. This step served as a procedural safeguard prior to piloting, rather than as an intervention to modify the AI-generated content. No substantive changes were made, thereby preserving the integrity of the AI-generated writing test prompts for performance evaluation. These writing test prompts, along with those developed by human item writers, were administered to a sample of 63 test-takers under standardized testing conditions consistent with CILS requirements. Test-takers are allotted 3 hours and 45 minutes for the full CILS B2 exam, which includes all components except for the speaking test. The written production and interaction component typically lasts 1 hour and 10 minutes, covering Task 1 and Task 2. However, in this study, only the written interaction and production section was administered. Since each task included two writing test prompts—one human-generated and one AI-generated—test-takers responded to four writing test prompts in total (two for Task 1 and two for Task 2). To account for the additional writing test prompts, the total time allotted for this session was 2 hours and 20 minutes.

Survey (Survey 2) responses from test-takers were analyzed to evaluate perceptions of

the writing test prompts. The survey explored familiarity with AI, experiences with the tasks, and perceived difficulty. Open-ended responses were further examined to identify themes, including sentiment toward the content and design of the prompts. These findings provided valuable insights into the extent to which the prompts adhered to the test takers' perspectives of expected characteristics of written production tasks in the CILS B2 exam. Therefore, test-takers' responses helped to investigate whether AI-generated prompts are perceived by test takers as appropriate and comparable to human-generated prompts, thereby supporting or challenging Assumption 3.

This study focuses solely on expert and test-taker perceptions, without analyzing test-taker performances. Future research could extend this work by examining test-taker responses to AI-generated versus human-crafted prompts, to further assess whether AI-generated prompts not only align with CILS task specifications but also successfully elicit B2-level writing performances.

Results

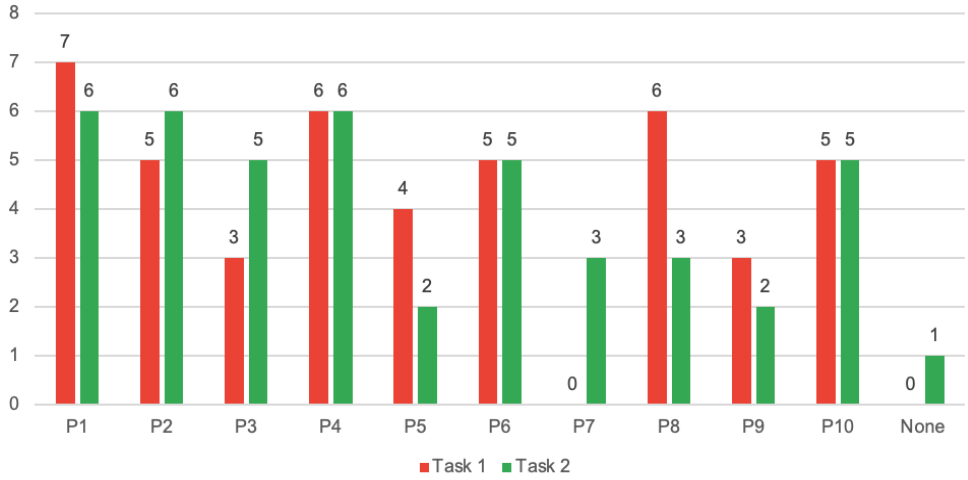
For RQ1, the results of Survey 1 were analyzed. Regarding the first question of Survey 1, the SMEs' responses are summarized in Figure 10.1, which also includes the content of the question in the caption. Figure 10.1 refers to the responses related to both the writing test prompts generated for Task 1 (no. 10) and those generated for Task 2 (no. 10).

For Task 1, Prompt 1 (P1) received unanimous approval from all SMEs, along with Prompts P4, P5, P8, and P10, which were also frequently identified as suitable. In contrast, P7 was not selected by any SME. For Task 2, P1, P2, P4, and P10 received the highest endorsements. However, there was greater variability for the remaining prompts. Prompt P5 and P9 were least endorsed, with only 2 experts selecting them. Notably, one SME deemed none of the writing test prompts appropriate for Task 2. Overall, across the 20 AI-generated writing test prompts, there was considerable variation in the acceptability among experts.

Regarding the open-ended question addressed to SMEs about the reasons behind the adequacy of each writing test prompt, their responses focused on the following aspects. For the prompts generated for Task 1, the SMEs considered the selected writing test prompts more or less relevant based on their suitability for assessing written production and their alignment with the B2 level descriptors of the CEFR (e.g., SME.6 "The writing test prompt is appropriately structured for a B2 level because the candidate is required to reflect on a topic, describe a personal experience in detail by explaining its advantages and disadvantages, and provide arguments."). Additionally, the SMEs reflected on the appropriateness of the content, the use of similar writing test prompts in previous CILS exam sessions with which they were familiar, and the use of specific verb forms (e.g., SME.5 "The test-takers is required to produce a text on a topic that may fall within their field of interest, using arguments and examples to support their ideas [B2]. The verb 'talk about' [ita. *parla*] is more suitable for a speaking task, so I would suggest changing it accordingly.").

Figure 10.1

Survey 1, Question 1: Number of SMEs Selecting 20 AI-generated Prompts as Suitable for Task 1 (Red) and Task 2 (Green) in the CILS B2 written production component (N=7)



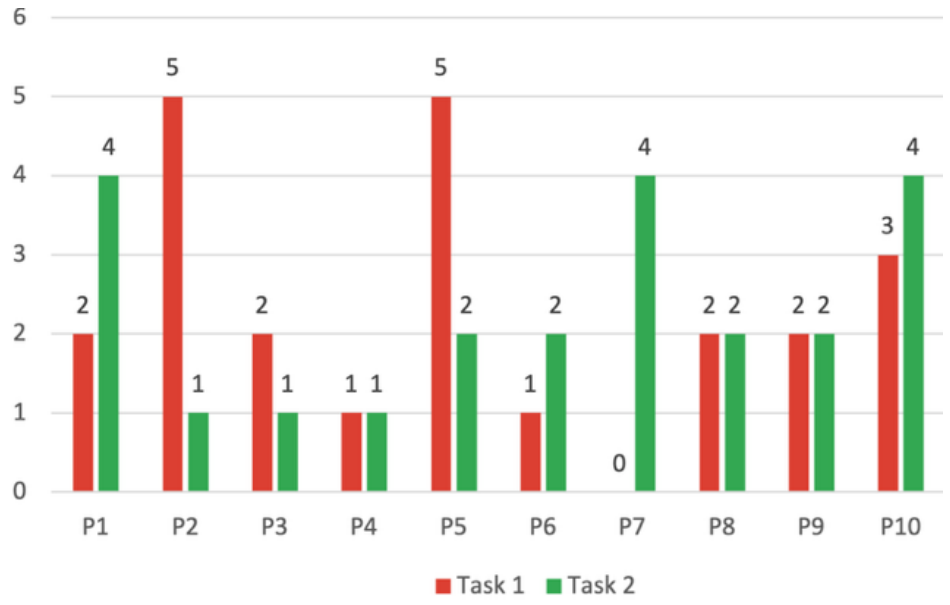
Note. P = prompt; None = no prompt is appropriate.

For the writing test prompts generated for Task 2, the SMEs primarily commented on their adequacy by considering morphosyntactic elements, the characteristics of the prompt formulation, and their relevance for written production and the complexity of the themes addressed (e.g., SME.4 "The topic should be related to the test-taker's field of interest at the B2 level"). Furthermore, for Task 2, the SMEs found it particularly useful to reflect on the interest and engagement generated by the prompts and mentioned the appropriateness of certain specific terms used in the prompts (e.g. SME.5 "I would remove the gerund—as it is not a structure typically expected at B2—and replace the term 'dean' [ita. *decano*] with a more general one.").

Regarding the second survey question on the use of AI for generating the writing test prompts, Figure 10.2 presents the collected data. As with the previous case, Figure 10.2 includes responses related to both the writing test prompts generated for Task 1 (no. 10) and those generated for Task 2 (no. 10).

Figure 10.2

Survey 1, Question 2: Number of SMEs Identifying 20 Writing Test Prompts as AI-generated (N=7)



Note. P = prompt

For Task 1, prompts P2 and P5 were most frequently suspected of being AI-generated, with 5 indications each. Some SMEs also identified P1, P3, P9, and P10 as potentially AI-generated, while P7 received no such indications. Notably, 2 SMEs indicated they believed none of the writing test prompts were AI-generated. For Task 2, prompts P1, P7, and P10 received the highest number of AI-generation suspicions, with 4 indications each. As with Task 1, 2 SMEs believed that none of the prompts were AI-generated.

The distribution of responses and accompanying comments suggest that there is no unanimous perception among experts about which prompts were AI-generated (e.g., SME.3 "There are no indicators that can distinguish whether the inputs were formulated by a human item writer or by artificial intelligence."). Also, this lack of agreement suggests the possibility that the AI-generated writing test prompts were not easily distinguishable from those created by humans. It is also plausible that their diversity was sufficient to approximate the variety typically expected in human-generated writing test prompts.

For RQ2, and after a human preliminary review process (see Data Collection and Analysis), the two best generated prompts—one for Task 1 and one for Task 2—were selected to be administered alongside the human-generated writing test prompts. Prompts 1 and 2 were the human-generated ones, while Prompts 3 and 4 were the AI-generated ones. However, this information was never revealed to test-takers. After the administration, test-takers responded to Survey 2. The following figures present the data collected.

The first question of the second survey investigated the perceived difficulty of the writing test prompts according to test-takers. The test takers' responses to this question are summarized in Figure 10.3 below.

Figure 10.3

Survey 2, Question 1: Histograms Showing Test Takers' Judgments of the Difficulty of Four Prompts (Prompt 1 and 2 Written by Humans; Prompts 3 and 4 AI-Generated) (N=63)

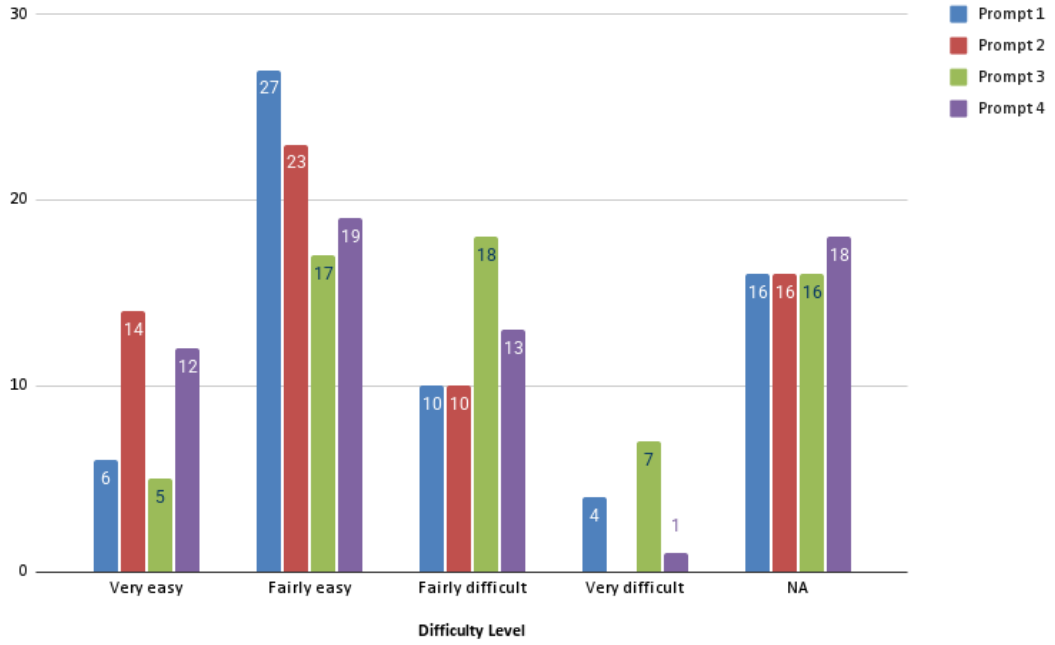


Figure 10.3 shows that "Fairly easy" has the highest number of responses in most cases, particularly in Prompts 1 and 2. The "Fairly difficult" rating is consistent but not dominant across prompts, with a spike in Prompt 3 (18 responses). In fact, there is a noticeable rise in both the "Fairly difficult" and "Very difficult" responses in Prompt 3, which indicate that this writing test prompt was perceived as more challenging compared to others. In contrast, Prompt 2 has higher "Very easy" and "Fairly easy" ratings, suggesting it was the least challenging overall.

A comparison between the human-generated prompts (1&2) and AI-generated ones (3&4) reveals differences in perceived difficulty, suggesting that AI-generated writing test prompts were generally perceived as more difficult. As shown in Figure 10.3, human-generated writing test prompts were more frequently rated as easier. For Prompt 1, 33 (70.2 % of valid responses) test takers rated it as "Very easy" or "Fairly easy", and for Prompt 2, 37 (78.7% of valid responses) test takers gave it these rating, making it the easiest overall. In contrast, the AI-generated writing test prompts, particularly Prompt 3, were perceived as more difficult. As mentioned before, Prompt 3 was judged as "Fairly difficult" or "Very difficult" by 25 (53.2 % of valid responses) test takers.

To better interpret test takers' perceptions of the writing test prompts, the second and third questions of Survey 2 investigated the familiarity of test takers with AI. Familiarity with AI

was considered a contextual factor to take into account, as individuals with greater exposure to AI might have different perceptions or expectation regarding the writing test prompts' characteristics. Furthermore, because the study investigated test takers perceptions of potential AI involvement in writing test prompt creation (Figure 10.6), understanding their baseline familiarity with AI was essential to interpret their judgements.

The data related to AI familiarity is presented in Figures 10.4 (Test-takers' knowledge of AI) and 10.5 (Test-takers' usage of AI) below.

Figure 10.4

Survey 2, Question 2: Test-takers' Knowledge of AI (N=63)

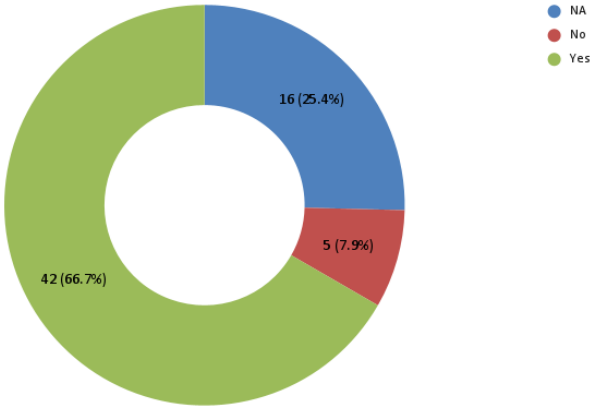


Figure 10.5

Survey 2, Question 3: Test-takers' Usage of AI (N=63)

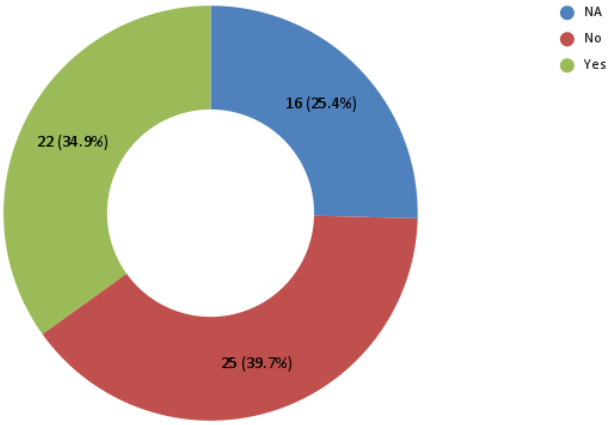


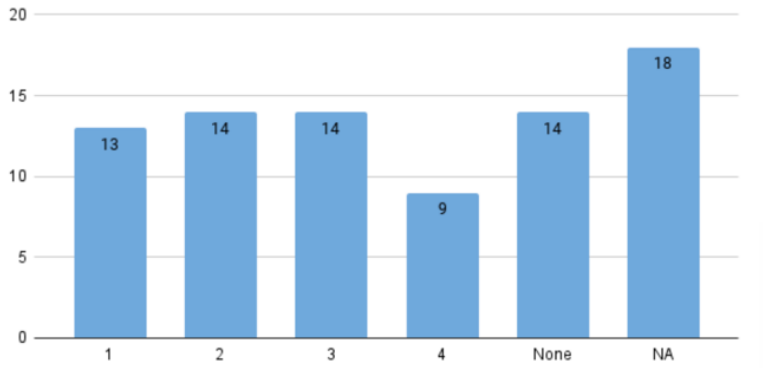
Figure 10.4 reveals that the majority of test-takers (66.2%) were familiar with the concept of AI. However, as shown in Figure 10.5, a comparable proportion of test-takers reported never having used AI (39.7%) and having used it in the past (34.9%).

For the fourth question of the second survey, test-takers were asked to give their opinion on the potential support of AI in creating the prompts they responded to. The data collected is

presented in Figure 10.6 below.

Figure 10.6

Survey 2, Question 4: Test-takers' Perception of AI Involvement in Writing Test Prompts (N=63)



Note. NA = the number of test-takers who did not respond.

As can be seen in Figure 6, Prompts 2 and 3 received the highest number of responses (14 each), indicating that test-takers were more likely to attribute these writing test prompts to AI (while only Prompt 3 was AI-generated) compared to Prompts 1 and 4 (while only Prompt 1 was human-generated). Interestingly, an equal number of test-takers (no. 14) believed that none of the prompts were created using artificial intelligence, reflecting a degree of skepticism or uncertainty regarding AI's role in content generation. Additionally, 18 test-takers did not engage with the question (NA), the highest count among all response categories. This may suggest that the question was unclear to some test-takers or that they lacked confidence in their ability to discern AI-generated content.

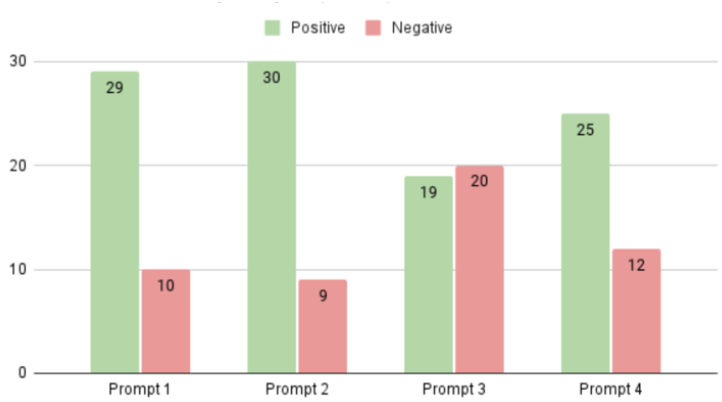
The final question of the second survey invited test-takers to reflect on their experiences with the writing test prompts in greater detail. Test-takers were asked to provide a brief description of their feelings while responding to each prompt. This included sharing whether they enjoyed the task, any difficulties they encountered, and whether they found the topic appropriate for their age group. The open-ended nature of this question aimed to capture nuanced insights into their engagement with the writing test prompts and their overall perceptions of the task. Test-takers' responses were grouped based on sentiment towards the content of the writing test prompt (Figure 10.7). Positive responses where test-takers expressed enjoyment, ease, or appropriateness of the topics were coded as positive (e.g., TT[2].3 "It was easy to write because I had already done something similar in my language course."). Negative responses, where test-takers mentioned difficulties, discomfort, or felt the topic was inappropriate for their age or context were coded as negative (e.g., TT.6 "The topic […] it's not very suitable for my age since we don't have much experience."). In two instances the test-takers' response was not clearly identifiable as positive or negative, as their language proficiency was probably too low to explain

[2] TT = Test taker

their point of view clearly (e.g., TT.9 “If I don’t know the language I want to use, I will have difficulty writing the email.”). These responses were not categorized.

Figure 10.7

Survey 2, Question 5: Sentiment Analysis of Test-takers Responses (N=39)



As shown in Figure 7, the analysis of sentiment toward the four writing test prompts reveals variations in test-taker perceptions. Prompt 2 received the most positive feedback, with 30 test-takers expressing favorable views and only 9 negative responses, suggesting it was the most well-received and effective writing test prompt. Similarly, Prompt 1 was positively regarded by the majority, with 29 positive responses and 10 negative ones, indicating strong overall approval. Prompt 4 also elicited predominantly positive feedback (25 positive responses), though it had slightly higher negative sentiment (12 responses) compared to Prompts 1 and 2. In contrast, Prompt 3 stood out as the most polarizing, with almost an equal split between positive (no. 19) and negative (no. 20) sentiments. This division may indicate challenges with its content, complexity, or alignment with expectations.

Discussion

The findings of this study provide empirical evidence on ChatGPT-4’s capability to generate writing test prompts that align with the CILS B2 exam, offering backing for the domain definition inference (Chapelle & Voss, 2021). Considering the judgment of SMEs, the study found an overlap between writing test prompts identified as AI-generated and those considered adequate for the exam, supporting Assumption 2 that AI-generated content can meet expert standards for B2-level writing prompts. However, expert feedback suggests areas for improvement, particularly by refining language complexity and ensuring alignment with the exam’s format, which could further refine the quality of AI-generated writing test prompts. This finding emphasizes the need for human oversight in AI-generated test materials.

Test-takers provided backing for Assumption 3 by expressing their judgment on the writing test prompts. The majority of test takers rated the writing test prompts as “Fairly easy”, with only one writing test prompt perceived as quite challenging. Although the comparison between the human-generated writing test prompts and AI-generated writing test prompts

revealed that the AI-generated ones were generally perceived as more difficult, this perception does not necessarily indicate inappropriateness, particularly when accompanied by positive sentiment (see below).

Interestingly, while 50% of the writing test prompts were commonly believed to be created by AI, 14 test-takers believed that none of the writing test prompts were created using AI, suggesting their uncertainty about AI's role in content generation. Despite this, test-takers had mainly a positive sentiment towards the writing test prompts.

These findings contribute to the validation of the domain definition inference, supporting the assumption that AI-generated prompts can be designed to align with the characteristics of the CILS B2 written interaction and production construct (RQ1) and are perceived as appropriate for eliciting B2-level writing skills (RQ2). However, full validation of the warrant requires further evidence, as this study does not yet evaluate test-taker written performances on AI-generated writing test prompts, as previously noted. The next stage of this research will focus on the evaluation inference (Chapelle & Voss, 2021), analyzing whether the written responses produced on AI-generated writing test prompts demonstrate the expected linguistic and strategic competencies aligned with CEFR B2 descriptors.

Conclusion

This study examined ChatGPT-4's capability to generate writing test prompts for the CILS B2 exam, providing backing for the domain definition inference. The findings suggest that AI-generated prompts can serve as viable alternatives to human-created ones, particularly in terms of clarity, thematic relevance, and perceived difficulty. However, refinements are needed in language complexity and adherence to CILS format, underscoring the continued need for human oversight in AI-assisted test development.

To further validate these findings, the next phase of this research will analyze test-taker written performances on AI- and human-generated prompts, offering additional backing for the domain definition inference while also contributing to the evaluation inference (Chapelle & Voss, 2021). This step will assess whether AI-generated prompts elicit responses that align with CEFR B2 descriptors.

Thus, this study contributes both practically and theoretically to AI-assisted language assessment. Practically, it explores how AI can enhance test development efficiency by reducing the time and effort required for writing test prompts. Theoretically, it advances understanding of the validity and appropriateness of AI-generated prompts in high-stakes language testing, particularly within S/FL Italian. These insights serve as a foundation for future research on AI's role in language test design, reinforcing the importance of validation frameworks in AI-driven assessment.

References

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, Article 903077. https://doi.org/10.3389/frai.2022.903077

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–217). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410602145-10

Centro CILS. (2009). *Linee guida CILS*. Guerra Edizioni. https://cils.unistrasi.it/public/articoli/12/linee_guida_cils_pdf(2).pdf

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Chapelle, C., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press. https://doi.org/10.1017/9781108669849

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. Routledge.

Cinganotto, L., Sbardella, T., & Montanucci, G. (2024). Dagli algoritmi alle competenze linguistiche: Il ruolo dell'intelligenza artificiale nell'educazione linguistica online. *The Journal of Language and Teaching Technology, 6*, 1-12. https://italian.rutgers.edu/images/PDFs/6.Cinganotto.Sbardella.Montanucci_JLTT_2024.pdf

Council of Europe (CoE). (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Gallina, F., Machetti, S., & Masillo, P. (2019). Il valutatore delle competenze linguistico-comunicative: Riflessioni e proposte per una figura professionale. In C. Bagna & V. Carbonara (Eds.), *Le lingue dei centri linguistici nelle sfide europee e internazionali: Formazione e mercato del lavoro* (pp. 67–83). Edizioni ETS.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Jun, H. (2021). Justifying the interpretation and use of an ESL writing final examination. In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 235–263). Cambridge University Press. https://doi.org/10.1017/9781108669849.014

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35(4)*, 477–499. https://doi.org/10.1177/0265532217716735

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, A. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Lee, E. (2023). L2 students' views on writing tools: Investigating domain definition within an argument-based validation framework. *English Teaching*, *78*(1), 125–144. https://doi.org/10.15858/engtea.78.1.202303.125

Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, *211*, Article 104977. https://doi.org/10.1016/j.compedu.2023.104977

Machetti, S. (2022). Le certificazioni di lingua italiana nei panorami linguistici globali: Spunti per un'analisi quantitativa. *Studi Italiani di Linguistica Teorica e Applicata, 51*(2), 452–476. https://perma.cc/8VH2-9KWY

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://openai.com

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the Eighth Conference on Machine Translation* (pp. 392–418). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.40

Rossi, O. (2023, May). *Using AI for test item generation: Opportunities and challenges* [Webinar]. EALTA Webinar Series. https://perma.cc/33N4-8QST

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341–358). De Gruyter Mouton. https://doi.org/10.1515/9781614513827-023

Voss, E. (2024). Artificial intelligence in language assessment. In A. Kunnan (Ed.), *The concise companion to language assessment* (2nd ed., pp. 112–125). Wiley-Blackwell.

Voss, E., Cushing, S. T., Ockey, G. J., & Yan, X. (2023). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and

practice. *Language Assessment Quarterly, 20*(4–5), 520–532. https://doi.org/10.1080/15434303.2023.2288256

Xi, X. (2023). Advancing language assessment with AI and ML–Leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly, 20*(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488

Appendix A

Table A1

CILS B2 Exam Level Description (CILS Guidelines, 2009, p. 29)

Competence profile	It is the level that certifies full autonomy in communicative competence in Italian as a foreign language. The test-taker is able to communicate effectively during a stay in Italy for study purposes and to manage interactions with the Italian language and culture, even for work-related reasons. Interaction with native speakers develops without excessive effort or tension. A test-taker at this level of competence can understand the main ideas of even complex texts that relate both to everyday life and to more abstract concepts. Oral and written production is communicatively effective, even if it contains some errors. This is the minimum level of competence required for access to the Italian university system, for completing a study cycle within a short-term mobility program for students, as well as for teachers and researchers. It is also necessary for benefiting from scholarships awarded by the Italian state and for undertaking an internship as part of a diploma course or within companies.
Syllabus of Morphosyntactic Structures	In addition to the structures covered in the previous levels, test-takers are required to understand and manage the following structures of the Italian language: Allocutive pronouns Indefinite pronouns and adjectives Combined pronouns Pronominal particles Conjugation of the active and reflexive forms of regular and irregular verbs, as well as modal verbs, in the following moods and tenses: • Present indicative • Present perfect indicative (passato prossimo) • Imperfect indicative • Past absolute indicative (passato remoto) • Pluperfect indicative (trapassato prossimo) • Simple and anterior future indicative • Present conditional • Past conditional • Present and imperfect subjunctive • Present and past infinitive • Imperative Passive form (only recognition) Impersonal verbs Most common adverbs of judgment and doubt. Simple sentence structure: Volitional clauses using the subjunctive, indicative, and infinitive.

	Complex sentence structure: Coordinated disjunctive, conclusive, and correlative clauses. Complex sentence structure: Subordinate clauses including subject, final, comparative, real-conditional, explicit concessive, explicit consecutive, and implicit temporal clauses.
Pragmatics and Language Use	The test-taker is able to express themselves confidently, clearly, and politely in either a formal or informal register, depending on the situation and the person involved. They can appropriately participate in discussions, effectively managing turns of speech. The test-taker is familiar with the social and communicative norms and rules typical of various everyday situations.
Vocabulary	The test-taker is able to navigate texts containing words from the Basic Vocabulary of the Italian language, as well as common vocabulary, up to a maximum of 7%. In oral and written production, they can use words from the core lexicon and also part of the high-frequency vocabulary.

Table A2

CILS DUE-B2 Written Production Component (CILS Guidelines, 2009, p. 31)

Profile	The test-taker is able to produce written texts using simple structures while conveying information clearly and effectively from a communicative perspective on familiar topics or subjects of personal interests (see also CEFR B2 writing production and interaction descriptors).
No. of tasks	2 Task 1: Descriptive/Narrative text Task 2: Structured text (email/letter)
Task types	Description of familiar people or places, storytelling involving given characters, diary-style narration of a trip, an event, or a particular episode; informal letter to relatives or friends to express gratitude, request information, or share experiences while highlighting key points.
Text length	Task 1: 100 to 120 words. Task 2: 50 to 80 words.
Test duration	1 hour and 10 minutes

Appendix B

Table B1

Administered Writing Test Tasks

Type of generation	Task 1	Task 2
Human	(Ita.) L'amicizia è una componente importante nella vita di tutti. Che cosa ne pensi? Parla delle tue esperienze positive e negative in proposito. (Eng.) Friendship is an important aspect of everyone's life. What do you think? Discuss your positive and negative experiences on the topic.	(Ita.) Sei indeciso se fare una vacanza-studio all'estero. Scrivi un'email a un tuo professore dove: - spieghi dove, quando e che tipo di esperienza vorresti fare; - chiedi la sua opinione ed eventuali suggerimenti. (Eng.) You are unsure about whether to take a study vacation abroad. Write an email to one of your professors where you: - Explain where, when, and what kind of experience you would like to have; - Ask for their opinion and any possible suggestions.
AI	(Ita.) Rifletti sull'importanza dello sport e del gioco di squadra nello sviluppo personale. Hai esperienze personali che evidenziano questi aspetti? (Eng.) Reflect on the importance of sports and teamwork in personal development. Do you have any personal experiences that highlight these aspects?	(Ita.) Vuoi migliorare le tue competenze linguistiche e decidi di iscriverti ad un corso di livello avanzato. Scrivi un'email ad una scuola di lingue e chiedi informazioni sul corso, sui requisiti di ammissione e sulla data di inizio. (Eng.) You want to improve your language skills and decide to enroll in an advanced-level course. Write an email to a language school asking for information about the course, admission requirements, and the start date.

About the Authors

Sabrina Machetti is the Director of the CILS (Certification of Italian as a Foreign Language) Centre, Associate Professor of Educational Linguistics. Her main research interests are focused on language testing and assessment, learning and teaching Italian as a foreign/second language, and language policies.

Giulia Peri earned her Ph.D. in Applied Linguistics at the University for Foreigners of Siena. Her research interests focus on S/FL Italian teaching and learning, language testing and assessment and technology. She is a Research Fellow at the CILS Centre (Certification of Italian as a Foreign Language) and the Project Coordinator for the L2 Italian scenario-based assessment project at the University for Foreigners of Siena.

Paola Masillo is Ph.D. in "Linguistics and Teaching Italian as second language", University for Foreigners of Siena. Her main research focuses on learning Italian as a foreign/second language, language assessment and language policies. She is currently working in the technical-scientific and data processing area at the CILS Centre of the University for Foreigners of Siena.

Author Note

All authors contributed equally to the conceptualization of the research. Giulia Peri was responsible for writing the Methods and Results sections. Sabrina Machetti wrote the Introduction. Paola Masillo wrote the Conclusions. The authors declare no conflicts of interest.

Chapter 11

The Use of ChatGPT for Reading Test Development and Validation: Two Case Studies

Hongli Li, Roula Aldib, and Chad Marchong
Georgia State University

This chapter explores the application of ChatGPT in reading test development and validation through two case studies. Study 1 evaluates ChatGPT's ability to estimate item difficulty and discrimination for a reading comprehension test without response data from test takers. Results show that ChatGPT estimated item difficulty with moderate agreement with empirical data, though it struggled with estimating item discrimination. Study 2 investigates ChatGPT's capacity to answer multiple-choice reading items without access to passages, revealing correct response rates that notably surpassed two human participants. These findings indicate ChatGPT's potential as a tool for facilitating reading test development and validation, particularly in providing early-stage validity evidence when pilot data are unavailable. Future studies should explore diverse test formats and refine AI methodologies to enhance their utility in measurement practice.

Introduction

ChatGPT is a chatbot powered by large language models (LLMs), which are designed to interpret user prompts and generate responses with near human fluency (Mead & Zhou, 2024; Zhou et al., 2023). Recent research has demonstrated ChatGPT's capability in designing reading assessment tasks. For instance, ChatGPT has been used for generating reading passages as well as corresponding multiple-choice items for reading assessments (Bezirhan & von Davier, 2023). It has also been recognized as capable of producing multiple-choice items comparable in quality to those created by human experts (Dueñas et al., 2023). Moreover, items generated by ChatGPT

Cite as: Li, H., Aldib, R., & Marchong, C. (2025). The use of ChatGPT for reading test development and validation: Two case studies. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 219-234). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.11
©2025 Li, Aldib, and Marchong. Published under a CC BY license.

have demonstrated satisfactory psychometric properties, supporting its utility in test development (Bhandari et al., 2024).

In traditional test development, the early stages often involve field testing, during which researchers collect pilot data to conduct item analysis, such as item difficulty, item discrimination, and test reliability. Based on the preliminary results, test items may be revised or discarded as necessary to improve the quality of the assessment. Test validation, as defined in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), is "the process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (p.11). This process requires gathering multiple types of validity evidence, including evidence related to test content, response processes, internal structure, relationships with other variables, and the consequences of testing. These steps are critical to ensure the test is appropriate for its intended purposes.

Despite the growing evidence of generative AI's ability to generate test items, its role in various stages of test development and validation remains understudied. Therefore, this chapter explores the application of generative AI—specifically, different versions of ChatGPT and the OpenAI API—for reading test development and validation. For simplicity, the term "ChatGPT" will be used to refer collectively to different versions of the ChatGPT platform and the OpenAI API models utilized in this study. Where distinctions are necessary, specific versions or configurations (e.g., ChatGPT 4o, OpenAI API [Temperature = 0]) will be explicitly identified.

To investigate ChatGPT's utility in this domain, two case studies are presented using the same reading comprehension test. Study 1 examines ChatGPT's ability to estimate item difficulty and discrimination without access to response data. Study 2 explores ChatGPT's performance in answering multiple-choice reading items without access to the associated passages. The findings from these studies are expected to offer valuable insights into how generative AI can streamline test development and validation.

Study 1

ChatGPT, a leading generative AI model introduced in November 2022, has gained widespread recognition for its capabilities in developing test items (Bezirhan & von Davier, 2023; Dueñas et al., 2023). However, its potential for evaluating item psychometric properties—such as difficulty and discrimination—during test development remains unclear. Traditionally, this process involves data collection and item analysis based on examinees' responses, a process which could be costly and time-consuming. The purpose of this study is to assess the effectiveness of ChatGPT in evaluating item properties of a reading comprehension test without response data.

Item difficulty and item discrimination are two important psychometric characteristics. According to classical test theory (CTT), item difficulty refers to the proportion of test takers who answer an item correctly, known as the *p-value*. Higher p-values indicate easier items, while lower p-values correspond to more difficult items. Item discrimination, on the other hand, measures an item's ability to differentiate between test-takers with varying levels of the ability

being assessed. Discrimination is often quantified through the difference in p-values between high-scoring and low-scoring test-takers. Alternatively, the item discrimination index can be represented by the point-biserial correlation between an item's performance (scored as 0 or 1) and the total test score (Suen, 1990).

Lumley et al. (2012) identified ten factors to predict PISA reading item difficulty, including the number of features and conditions to be comprehended, proximity of pieces of required information, competing information, prominence of necessary textual information, relationship between task and required information, semantic match between task and text, concreteness of information, familiarity of information needed to answer the question, register of the text, and extent to which information from outside of the text is required to answer the question. While research on factors influencing item discrimination is rather limited, Crocker and Algina (2006) observed that items that are either too easy or too difficult tend to have lower discrimination power because they fail to effectively distinguish between high- and low-ability test-takers; items with moderate difficulty generally exhibit higher discrimination power, as they produce a balanced distribution of correct and incorrect responses; additionally, poorly constructed distractors—such as those that are implausible or overly obvious—reduce item discrimination.

Methods

The established procedure for estimating item difficulty and item discrimination involves analyzing response data collected from examinees. In this study, however, we investigated whether ChatGPT can estimate item difficulty and discrimination without access to response data. The reading comprehension test used in this study is a reading section of an English test, which is designed to evaluate advanced level English language competence of adult non-native speakers of English who plan to use English for academic purposes in a university setting. This test assesses examinees' understanding of college-level reading texts and includes four passages, each accompanied by five multiple-choice items, totaling 20 items. Each item has four options, including one answer key and three distractors. All four passages were adapted from newspaper articles.

Different versions of ChatGPT were asked to estimate the difficulty and discrimination of the 20 items without access to response data in October 2024. The versions tested included ChatGPT 4o, ChatGPT 4o mini (free version), OpenAI API (temperature=0), and OpenAI API (temperature=0.5). ChatGPT 4o is a paid version of ChatGPT offering enhanced performance and reliability, while ChatGPT 4o mini is a free, scaled-down version of ChatGPT 4o with limited capabilities compared to its paid counterpart. The temperature parameter in the OpenAI API controls the randomness of responses. It ranges from 0 to 2, with higher values producing more varied outputs. OpenAI API (Temperature = 0) uses a deterministic setting of the OpenAI API, producing consistent responses by minimizing randomness in output generation. OpenAI API (Temperature = 0.5) uses a slightly randomized setting of the OpenAI API, allowing for more varied responses. Comparing these two settings allowed us to evaluate whether introducing controlled randomness would enhance or diminish the AI model's accuracy or estimation capability in the specific context. Below are the sample prompts used to interact with the different versions of ChatGPT and the OpenAI API:

Researcher: Do you understand the concept of item difficulty and item discrimination?

ChatGPT: Yes, I do.

Researcher: Great! I will provide a reading section of a test. This test is designed to evaluate advanced level English language competence of adult non-native speakers of English who will use English for academic purposes in a university setting. The reading section of this test is designed to assess examinees' understanding of college-level reading texts. It consists of four passages, each of which is followed by five multiple-choice items. According to the item-writing guidelines, the questions following each passage are intended to assess a variety of reading abilities, including recognizing the main idea, understanding the relationships between sentences and portions of the text, drawing text-based inferences, synthesizing, understanding the author's purpose or attitude, and recognizing vocabulary in context. If I provide you with the reading passages and items, can you provide an estimate of the difficulty and discrimination of each item? I understand that you won't be able to get the exact item difficulty and discrimination without the item response. I only ask you to provide an estimation.

ChatGPT: Yes, I can provide an estimation. (Here, ChatGPT provided a rough description of whether the item difficulty and discrimination are low, medium, or high.)

Researcher: If you consider the framework of CTT item difficulty and item discrimination, can you provide more nuanced numerical values?

ChatGPT: Yes, I can.

To evaluate ChatGPT's estimates, item responses from 1,968 test takers were analyzed using SPSS, and these empirical results served as a benchmark for comparison.

Results

Table 11.1 and Figure 11.1 display item difficulty estimates generated by different AI models alongside the empirical estimates derived from CTT analysis. Table 11.2 presents the inter-correlations among these five sets of estimates. The correlation between ChatGPT 4o estimates and the empirical values was 0.484 ($p < .05$), between OpenAI API (Temperature=0) estimates and the empirical values was 0.490 ($p < .05$), between OpenAI API (Temperature=0.5) estimates and the empirical value was 0.451 ($p < .05$). According to Cohen (1988), .010, 0.30, and 0.50 were recommended to be considered small, medium, and large in magnitude, respectively. Therefore, our findings indicate that both ChatGPT 4o and the OpenAI API provided moderate correlations with the empirical values derived from the CTT analysis, suggesting reasonable performance in estimating item difficulty.

Table 11.1

Comparison of Item Difficulty Estimates between CTT Empirical Results and AI Models

Item Number	CTT Difficulty	ChatGPT 4o	ChatGPT 4o mini	OpenAI API (Temperature= 0)	OpenAI API (Temperature= 0.5)
1	0.55	0.50	0.55	0.60	0.65
2	0.65	0.80	0.60	0.80	0.75
3	0.72	0.40	0.50	0.50	0.55
4	0.76	0.50	0.55	0.60	0.60
5	0.36	0.30	0.60	0.50	0.55
6	0.47	0.70	0.60	0.60	0.65
7	0.74	0.50	0.55	0.70	0.70
8	0.44	0.40	0.50	0.50	0.55
9	0.81	0.80	0.75	0.80	0.75
10	0.27	0.30	0.50	0.50	0.55
11	0.62	0.50	0.55	0.60	0.65
12	0.60	0.80	0.60	0.70	0.70
13	0.75	0.60	0.60	0.60	0.65
14	0.54	0.50	0.55	0.50	0.55
15	0.44	0.30	0.50	0.50	0.55
16	0.45	0.50	0.55	0.60	0.65
17	0.32	0.60	0.60	0.60	0.65
18	0.51	0.50	0.70	0.80	0.75
19	0.45	0.40	0.55	0.50	0.55
20	0.56	0.50	0.65	0.60	0.65
Mean	.55	.52	.58	.61	.63
SD	.15	.16	.07	.11	.07

Figure 11.1

Scatterplot of Item Difficulty Estimates between CTT Empirical Results and AI Models

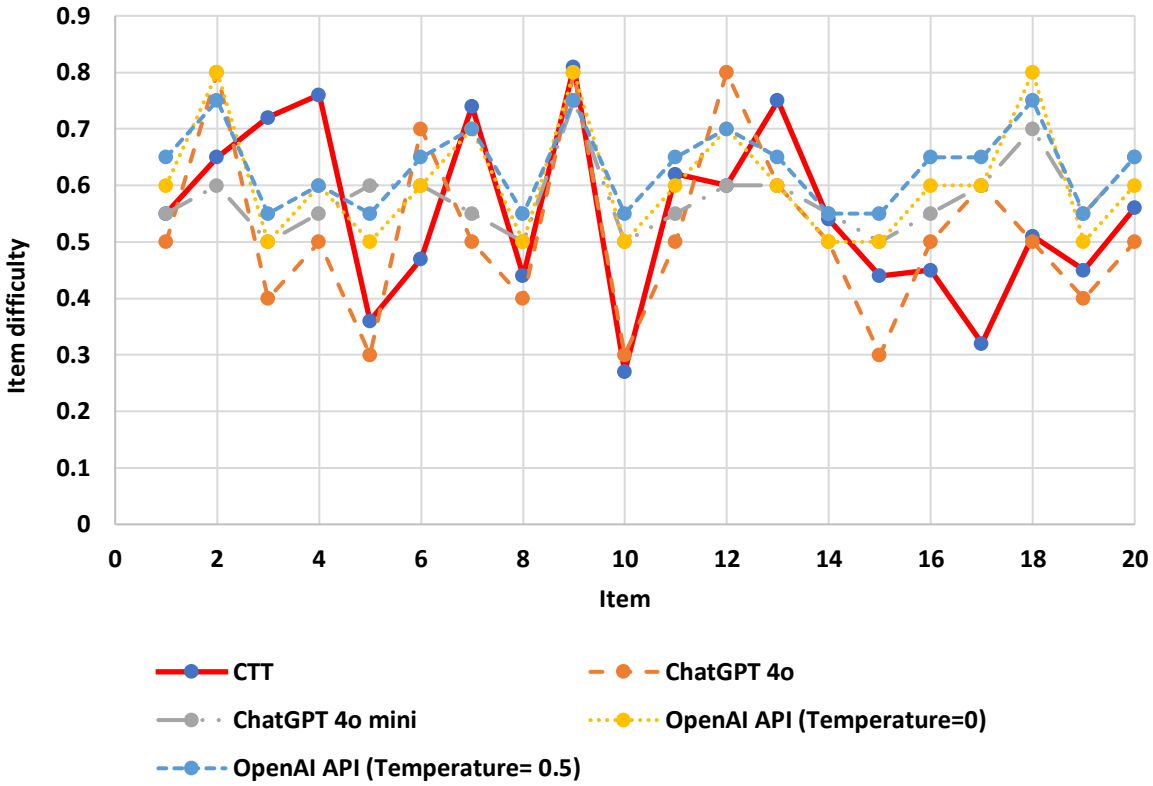


Table 11.2

Correlations among Item Difficulty Estimates across CTT Empirical Results and AI Models

	CTT Difficulty	ChatGPT 4o	ChatGPT 4o mini	OpenAI API (Temperature= 0)	OpenAI API (Temperature= 0.5)
CTT Difficulty	1	.484*	0.288	.490*	.451*
ChatGPT 4o	.484*	1	.604**	.757**	.786**
ChatGPT 4o mini	0.288	.604**	1	.740**	.735**
OpenAI API (Temperature= 0)	.490*	.757**	.740**	1	.973**
OpenAI API (Temperature= 0.5)	.451*	.786**	.735**	.973**	1

Note. *p < .05; **p < .01

Table 11.3 and Figure 11.2 present item discrimination estimates generated by different AI models alongside the empirical estimates derived from CTT analysis, while Table 11.4 presents the inter-correlations among these five sets of estimates. The correlations between the item discrimination estimations from all AI models and the empirical discrimination values were not statistically significant. Notably, except for the correlation between CTT estimate and ChatGPT 4o mini ($r = .352$, $p > .05$), all other correlations were close to zero. This suggests that ChatGPT struggles with estimating item discrimination, which is understandable given the inherent complexity of this task. Additionally, the empirical discrimination values of the 20 items showed little variation, which likely contributed to the small correlations due to the restricted range of values.

Table 11.3

Comparison of Item Discrimination Estimates between CTT Empirical Results and AI Models

Item Number	CTT Discrimination	ChatGPT 4o	ChatGPT 4o mini	OpenAI API (Temperature= 0)	OpenAI API (Temperature= 0.5)
1	0.31	0.40	0.45	0.40	0.35
2	0.41	0.30	0.50	0.30	0.30
3	0.36	0.50	0.35	0.50	0.45
4	0.41	0.40	0.30	0.40	0.40
5	0.36	0.50	0.30	0.50	0.45
6	0.34	0.30	0.50	0.40	0.35
7	0.38	0.40	0.45	0.30	0.30
8	0.42	0.45	0.35	0.50	0.45
9	0.33	0.25	0.30	0.30	0.30
10	0.35	0.50	0.40	0.50	0.45
11	0.34	0.40	0.50	0.40	0.35
12	0.37	0.30	0.35	0.30	0.30
13	0.41	0.35	0.50	0.40	0.35
14	0.47	0.45	0.40	0.50	0.45
15	0.24	0.50	0.30	0.50	0.45
16	0.45	0.40	0.50	0.40	0.35
17	0.26	0.35	0.40	0.40	0.35

Table 11.3 (continued)

Comparison of Item Discrimination Estimates between CTT Empirical Results and AI Models

18	0.42	0.40	0.35	0.30	0.30
19	0.45	0.45	0.50	0.50	0.45
20	0.44	0.40	0.55	0.40	0.35
Mean	.37	.40	.41	.41	.38
SD	.06	.07	.08	.08	.06

Figure 11.2

Scatterplot of Item Discrimination Estimates between CTT Empirical Results and AI Models

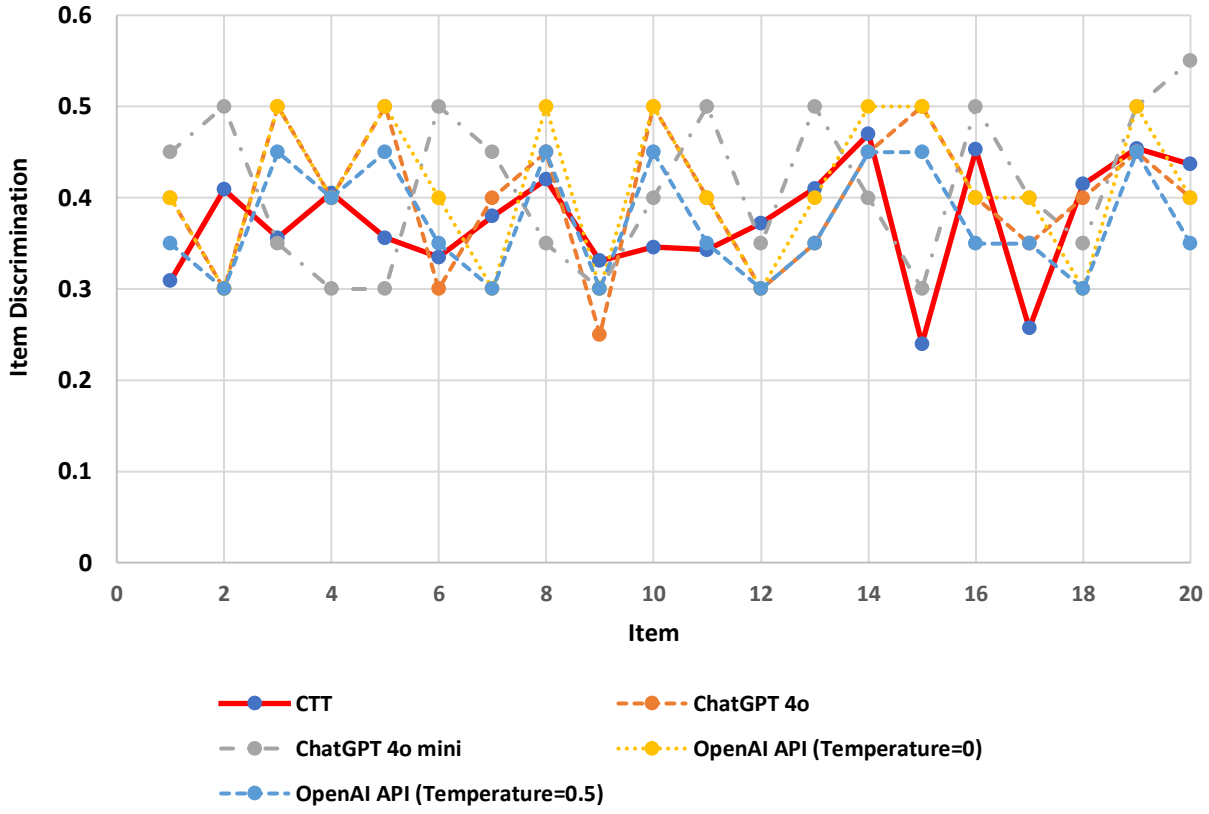


Table 11.4

Correlations among Item Discrimination Estimates across CTT Empirical Results and AI Models

	CTT discrimination	ChatGPT 4o	ChatGPT 4o mini	OpenAI API (Temperature= 0)	OpenAI API (Temperature= 0.5)
CTT discrimination	1	0.027	0.352	-0.016	0.022
ChatGPT 4o	0.027	1	-0.232	.809**	.831**
ChatGPT 4o mini	0.352	-0.232	1	-0.139	-0.291
OpenAI API (Temperature= 0)	-0.016	.809**	-0.139	1	.973**
OpenAI API (Temperature= 0.5)	0.022	.831**	-0.291	.973**	1

Note. *p < .05; **p < .01

Discussion

Our findings indicate that the estimates by ChatGPT 4o and the OpenAI API demonstrated moderate correlations with empirical item difficulty estimates, suggesting promising potential for AI-assisted item difficulty estimation. However, ChatGPT struggled with estimating item discrimination, as the correlations with empirical values were negligible. The results of this study are aligned with the findings reported by Pereira et al. (2024) in their analysis of a medical exam. In their study conducted between April and May 2023, ChatGPT 4 was queried 100 times with 150 multiple-choice items from the Portuguese National Residency Selection Exam (PNA). Using ChatGPT's simulated responses, item difficulty indices were calculated for each question. The predicted difficulty levels from ChatGPT were compared to the actual difficulty levels based on response data. There was a significant correlation between ChatGPT's predicted difficulty indexes and test takers' actual values ($r = .372$, $p < .001$). However, the correlation between the item discrimination estimation from the ChatGPT and the actual value was close to zero ($r = -.0257$, $p > .05$). These findings, coupled with our results, demonstrate the potential of generative AI in estimating item difficulty without access to response data in the early stages of test development.

Nevertheless, a key limitation of our study is the lack of direct comparison between AI-generated item difficulty estimates and human predictions. While estimates by ChatGPT and OpenAI API demonstrated moderate correlations with empirical item difficulty values, it remains unclear how these estimates compare to those made by human experts. Future research should explore this comparison to better contextualize the accuracy and practical utility of AI-based difficulty estimation.

Both our findings and those of Pereira et al. revealed that ChatGPT struggles with estimating item discrimination. This limitation may stem from the inherent complexity of discrimination as a psychometric property, which relies on subtle patterns in response data that are challenging to infer without sufficient empirical observations. For instance, many large-scale assessments use Rasch modeling to scale the items, focusing only on item difficulty. Also, when the discrimination values of items have minimal variations, it will be more difficult for AI tools to estimate item discrimination. Given this, the challenge observed in our study is not unique to ChatGPT but rather reflects the statistical nature of item discrimination itself. In that regard, estimating item discrimination without response data remains a significant challenge and may have inherent limitations.

Study 2: Responding to Multiple-Choice Items Without Reading Passages

Many reading comprehension tests use a multiple-choice format (Hartung et al., 2017) because this format is easy to implement and score (Keenan & Betjemann, 2006). Nevertheless, studies have raised concerns regarding the use of a multiple-choice format for the assessment of reading comprehension. Mainly, the problem lies in participants' ability to answer the multiple-choice questions correctly without accessing the associated reading passages at rates significantly above chance (Coleman et al., 2010; Daneman & Hannon, 2001; Katz et al., 1990). If test items can be answered correctly without referring to the passages, the test fails to assess passage comprehension, which is the intended construct. Instead, the test might assess other abilities (e.g., prior knowledge or reasoning; Katz et al., 1990). This phenomenon, referred to as passage-independent responding, poses a threat to the construct validity of reading comprehension tests because the scores may not accurately reflect examinees' comprehension abilities (Coleman et al., 2010). Given this concern, it is critical to investigate the construct validity of reading comprehension tests from this perspective; however, traditional test validation processes require substantial costs for test administration and data collection.

Nevertheless, the rise of large language models (LLMs), such as ChatGPT, offers a promising opportunity to address such challenges. LLMs are trained on extensive collections of text, enabling them to generate responses that mimic human reasoning and language use. ChatGPT uses advanced natural language processing (NLP) techniques to simulate comprehension and may therefore be a useful tool for assessing reading comprehension items. We chose ChatGPT over other generative AI models, such as Copilot or Gemini, because of its demonstrated performance in language-based tasks. For example, de Winter (2024) investigated ChatGPT's performance in generating responses to English reading comprehension exams in the Netherlands. The study found that ChatGPT's capabilities improved with subsequent iterations (i.e., while ChatGPT 3.5 scored similarly to Dutch students on the reading comprehension exam, ChatGPT 4 demonstrated higher performance on the same exam). Although de Winter's study highlights ChatGPT's versatility, no research has yet examined the tool's ability to generate passage-independent responses (i.e., to answer reading comprehension items without access to the reading passage).

The purpose of Study 2 is to examine how well ChatGPT can provide responses to multiple-choice reading items without access to the passages. We hypothesized that, given the

expansive background knowledge and analytical capabilities of ChatGPT, its correct response rate would represent the upper boundary of a typical test taker's ability to respond to such items without accessing the passages.

Methods

In this study, ChatGPT provided responses to 20 multiple-choice reading comprehension items without accessing the associated passages. The reading items were taken from an advanced-level English-language competence test for non-native English speakers, which was introduced in Study 1. We employed multiple ChatGPT models, including GPT 3.5, GPT 4, GPT4o, GPT4o mini, and the OpenAI API. ChatGPT 3.5 is a foundational version of OpenAI's generative AI models that served as the precursor to the more advanced GPT 4. We also used the OpenAI API version (with the temperature set as 0 and 0.5).

We used the following prompt to instruct ChatGPT to perform the task:

> *Researcher: Hello, here are 20 multiple-choice questions from a reading test. Question 1-5 are from passage 1. Questions 6-10 are from Passage 2. Questions 11-15 are from Passage 3, and Questions 16-20 are from Passage 4. Can you provide a correct answer to each question based on the provided questions only (without passages available)? Can you also provide your justifications?"*
>
> *ChatGPT: I'll start by examining the file you uploaded to understand the questions. Then, I'll provide the correct answers and justifications based on the questions alone. Let's take a look at the document first. Based on the questions and the choices provided, here are the most plausible answers, along with justifications for each.*

In addition, passage-independent responses were also obtained from two graduate student participants, and results were compared to those generated by the different AI models. One participant was a male native English speaker from the College of Education, and another participant was a female non-native English speaker from the Department of Psychology. Both participants were given the following instructions:

> *"Please select the correct answer for each of the following multiple-choice items on your own. Also, please write a brief justification of why you chose a particular answer. You can take as much time as needed."*

Results

As shown in Table 11.5, all versions of ChatGPT outperformed the two human participants, with correct rates ranging from 75% to 85%, compared to only 40% for humans. Notably, the paid ChatGPT versions (4o) and OpenAI API did not perform better than the free ChatGPT versions (3.5, 4o mini). Additionally, the OpenAI API produced the same result whether the temperature was set as 0 or 0.5. Also, ChatGPT's performance remained stable over time (from April to October 2024). This indicated that a hypothetical individual as "intelligent" as ChatGPT could achieve a high score on this reading comprehension test without reading the passages. Among the 20 items, both human participants and all versions of ChatGPT provided correct answers to

question 3. Upon examination, question 3 was found to be an easy item, as the question required general knowledge that one may already know without reading the passage.

When asked about the rationales used to provide the responses, ChatGPT reported relying on the following resources: reasoning (i.e., typical question structures, common themes in reading comprehension tests, patterns in standardized reading comprehension tests, logical associations, test-taking strategies), and background knowledge (i.e., familiarity with common topics, understanding of scientific and social concepts). The two graduate students reported using similar strategies. Also, both ChatGPT and the graduate students mentioned using educated guesses when selecting their answers. ChatGPT provided answers and justifications instantly, while both graduate students spent about one hour on the whole task and reported feeling frustrated while responding to the multiple-choice items without access to the passages.

Table 11.5

Summary of Item Responses

Item	Person 1	Person 2	ChatGPT 3.5 (04/18/)	ChatGPT 4 (04/18)	ChatGPT 4o (09/18)	ChatGPT 4o mini (09/18)	OpenAI API (10/04)
1	1	1	0	0	1	1	1
2	0	0	1	1	1	0	1
3	1	1	1	1	1	1	1
4	0	0	1	1	1	1	1
5	1	1	1	1	1	0	1
6	0	0	1	0	0	0	0
7	1	1	1	1	0	1	0
8	0	0	1	1	1	1	1
9	0	0	1	1	1	1	1
10	0	1	1	1	1	1	1
11	0	1	1	1	1	1	1
12	1	0	1	0	0	0	0
13	0	0	0	0	1	1	1
14	0	0	1	1	1	1	1

Table 11.5 (continued)

Summary of Item Responses

15	0	0	1	1	1	1	1
16	1	0	1	0	1	1	1
17	1	0	1	1	1	1	1
18	0	0	1	1	1	1	1
19	1	1	0	1	1	1	1
20	0	1	1	1	1	1	1
Proportion of correct	40%	40%	85%	75%	85%	80%	85%

Note. 1 indicates correct, and 0 indicates incorrect.

Discussion

Concerns have been raised regarding the construct validity of reading comprehension tests, particularly when students can correctly answer multiple-choice items without thoroughly engaging with the accompanying passages (Keenan & Betjemann, 2006). In this study, the two graduate students achieved a correct response rate of 40%, which is notably higher than the chance correct rate of 25% for typical multiple-choice items. Our pilot findings are aligned with those reported in Coleman et al. (2010). In their study, Coleman et al. invited 253 typical undergraduate students and 26 at-risk undergraduate students to respond to the Nelson-Denny Reading Test (NDRT). They found that typical students had correct rates of 43.6% and 46.6%, while at-risk students had a correct rate of 40.6%. The authors concluded that the NDRT exhibited significant issues related to passage independent responding. All four reading passages in the current study were adapted from newspaper articles, primarily consisting of expository essays. Responses to questions about such passages may be particularly susceptible to passage independence, as examinees might rely on their background knowledge to answer questions correctly without engaging deeply with the text.

Examining the construct validity of reading comprehension tests through traditional piloting methods, as outlined by Coleman et al. (2010), requires extensive resources for test administration and data collection. In contrast, this study demonstrated that ChatGPT offers a cost-effective and efficient alternative for examining passage independence in reading comprehension tests, providing evidence relevant to construct validation. Furthermore, ChatGPT can provide detailed rationales for its answers. These rationales may inform test revisions by identifying potential weaknesses or flaws in item writing.

Conclusion

This chapter explored the application of ChatGPT in reading test development and validation through two case studies. In Study 1, ChatGPT demonstrated the ability to estimate item difficulty with moderate agreement with empirical values, highlighting its potential for supporting early-stage test development. Estimating item discrimination, however, proved more challenging, likely due to the inherent complexity of discrimination metrics and the restricted range of empirical values in the dataset. Study 2 further extended the investigation by examining ChatGPT's ability to answer multiple-choice reading comprehension items without access to the accompanying passages. The AI tool's high correct response rates highlight its potential utility as an upper-bound benchmark for evaluating passage independence. In this way, ChatGPT could provide important validity evidence, identifying items that may be answered correctly through reasoning or background knowledge rather than passage comprehension. Moreover, its ability to generate detailed rationales can help identify item design flaws.

In conclusion, ChatGPT represents a valuable tool that complements traditional methods in reading test development and validation. However, the two studies reported here are exploratory in nature, and a major limitation is that only a single reading test was used. Notably, the test examined in this study was used for demonstration purposes only, as the actual test had already been discontinued. This lowers the stakes of our findings but allows for an initial exploration of ChatGPT's capabilities in these tasks. Additionally, the prompts used to instruct ChatGPT in this study were relatively simple. More detailed prompts and targeted training could further improve the accuracy and informativeness of ChatGPT's estimates. Future studies should apply ChatGPT to various test formats, refine its use in test development and validation, and explore AI training strategies to improve its estimation precision.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5,* 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education: Artificial Intelligence, 7*, Article 100284. https://doi.org/10.1016/j.caeai.2024.100284

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg, K. N. (2010). Passageless comprehension on the Nelson-Denny reading test: Well above chance for university students. *Journal of Learning Disabilities, 43*(3), 244–249. https://doi.org/10.1177/0022219409345017

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Wadsworth.

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General*, *130*(2), 208-223. https://doi.org/10.1037//0096-3445.130.2.208

de Winter, J. C. F. (2024). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education, 34*, 915–930. https://doi.org/10.1007/s40593-023-00372-z

Dueñas, G., Jimenez, S., & Gerro, G. (2023). You've got a friend in . . . a language model? A comparison of explanations of multiple-choice items of reading comprehension between ChatGPT and humans. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, (pp. 372–381). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.30

Hartung, J., Weiss, S., & Wilhelm, O. (2017). Individual differences in performance on comprehension and knowledge tests with and without passages and questions. *Learning and Individual Differences*, *56*, 143–150. https://doi.org/10.1016/j.lindif.2016.11.001

Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science, 1*(2), 122–127. https://doi.org/10.1111/j.1467-9280.1990.tb00080.x

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, *10*(4), 363–380. https://doi.org/10.1207/s1532799xssr1004_2

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012). *A framework for predicting item difficulty in reading tests*. Australian Council for Educational Research (ACER). https://research.acer.edu.au/pisa/5/

Mead, A. D., & Zhou, C. (2024). Evaluating the quality of AI-generated items for a certification exam. *Journal of Applied Testing Technology*, *24*, 1–53. https://jattjournal.net/index.php/atp/article/view/173204

Pereira, D. S. M., Mourão, F., Ribeiro, J. C., Costa, P., Guimarães, S., & Pêgo, J. M. (2024). ChatGPT as an item calibration tool: Psychometric insights in a high-stakes examination. *Medical Teacher, 47*(4), 677–683. https://doi.org/10.1080/0142159X.2024.2376205

Suen, H. K. (1990). *Principles of test theories*. Lawrence Erlbaum Associates.

Zhou, T., Zhang, Y., Cao, S., Li, Y., & Wang, T. (2023). *Complementary advantages of ChatGPT and human readers in reasoning: Evidence from English text reading comprehension*. ArXiv. https://doi.org/10.48550/arXiv.2311.10344

About the Authors

Hongli Li is a professor in the Program of Research, Measurement, and Statistics (RMS) at the Department of Educational Policy Studies, Georgia State University. Her primary research areas are applied measurement, language assessment, and quantitative methods. In particular, she is interested in test validity and fairness, and how testing influences teaching and learning (cognitive diagnostic assessment and formative assessment). Her most recent work includes peer assessment and the use of AI in test development and validation. Here is her website: https://education.gsu.edu/profile/hongli-li/

Roula Aldib is a Ph.D. student in the Developmental Psychology Program at Georgia State University, where she also earned her master's degree. Her research centers on adult literacy and numeracy, with a particular interest in the measurement of these abilities. By leveraging generative AI, Item Response Theory (IRT), and other psychometric methods, she aims to develop equitable assessments for diverse populations. Her experience includes validating assessments using measurement models, such as IRT and CFA, and analyzing data from large-scale studies on adult cognitive skills.

Chad Marchong is a Ph.D. student in Educational Policy Studies at Georgia State University, concentrating on Research, Measurement, and Statistics. His research interests include using item response theory (IRT) and differential item functioning (DIF) to understand the impact of students of diverse backgrounds' performance on large-scale standardized assessments. Additionally, he is interested in the possibilities of item development, IRT, and generative AI. Chad has over 15 years of experience working in educational environments, including K-12 and higher education. He is the Associate Director of Learning Analytics and Research at Georgia State University.

Chapter 12

Exploring ChatGPT's Ability to Recognize and Generate Language Variations: North vs. South Korean Dialects

Jean Young Chun
Defense Language Institute Foreign Language Center (DLIFLC)

This study explored the competence of ChatGPT-4o (OpenAI, 2023) in recognizing and producing non-English dialects, specifically South Korean (SK) and North Korean (NK), aiming to support the development of dialect teaching and testing materials—a largely underexplored area. A total of 100 SK passages and 100 NK passages were used to evaluate its performance in recognition tasks, including dialect distinction, identification of NK words and their meanings, and production tasks, such as translating the SK dialect into the NK dialect. The study found that ChatGPT-4o was highly capable of recognizing each target dialect and identifying NK words and their meanings. However, its performance declined noticeably when converting the SK dialect to the NK dialect. These findings reveal a significant gap in its ability to handle dialects, shedding light on both its potential applications and limitations in language teaching and testing.

Introduction

The advancement of AI technology holds great promise for the development of language teaching and testing materials. ChatGPT (Generative Pre-trained Transformer), which is one of the Large Language Models (Aryadoust et al. 2024), has been rapidly integrated into language learning due to its accessibility and ease of use by language practitioners, such as teachers and test developers. This highly advanced large language model simulates human-like conversations and responds promptly to user requests, driving its widespread adoption and establishing its role in the field. Among its various contributions, ChatGPT's text generation has garnered increasing

Cite as: Chun, J. Y. (2025). Exploring ChatGPT's ability to recognize and generate language variations: North vs. South Korean dialects. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 235-251). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.12
©2025 Chun. Published under a CC BY license.

attention as a tool capable of producing diverse texts tailored to users' needs (Kohnke et al., 2023).

While ChatGPT offers this convenience, it is essential to evaluate whether ChatGPT-generated texts accurately represent the language use of the target domain and whether they can serve as a valid resource for language learners. This question addresses the concept of authenticity, which is one of the key factors in determining the quality of testing materials. According to Bachman (1990), authenticity refers to "the extent to which test tasks replicate real-life language use tasks" (p. 307). In other words, the linguistic similarity between test texts —an integral component of test tasks—and real-life language use enables a more accurate interpretation of learners' ability to handle texts beyond the testing context. Thus, to ensure authentic input for language learners, empirical scrutiny of ChatGPT's texts is necessary to assess their representativeness of language use within the target domain, which should precede their integration into language instruction or assessment.

The efforts to validate the quality of texts have been well-documented in English assessments, demonstrating the great potential of leveraging ChatGPT to develop English texts. Xiao et al. (2023) found that ChatGPT-generated English reading texts exhibited greater lexical diversity than texts generated by the GPT-2 model and human-written passages. Additionally, ChatGPT passages received higher scores from human evaluators on a range of criteria, such as coherence, correctness, and engagement, than the other two types of passages and were rated as more human-like than texts created by human writers. Supporting these findings, Shin and Lee (2023) reported that ChatGPT-generated English reading passages were comparable to those written by humans. ChatGPT passages were rated by pre-service and in-service teachers as highly as those by human authors. Beyond reading texts, ChatGPT has also shown promise in developing listening scripts. Aryadoust et al. (2024) used ChatGPT-4 to generate English academic listening scripts—primarily lectures—targeting different difficulty levels (low, intermediate, and advanced). Their findings revealed that ChatGPT-4 generated scripts aligned with the expected difficulty levels and showed significant differences in readability, lexical diversity, syntactic complexity, and the use of academic vocabulary across levels.

These studies suggest that ChatGPT can serve as a reliable resource for producing high-quality English texts. However, its ability to contribute comparably to the development of passages in other languages remains uncertain. Lai et al. (2023) found that ChatGPT's accuracy in multiple Natural Language Processing (NLP) tasks, such as commonsense reasoning and summarization, varied across languages and did not align closely with the level of available language resources, ranging from high-resource to low-resource languages. According to Zhang et al. (2023), ChatGPT, primarily trained on English datasets, tends to produce outputs resembling English, reducing its accuracy in other languages, particularly those with non-Latin scripts. They argue that ChatGPT adopts a subordinate bilingualism model, relying heavily on literal translations between English and target languages, often resulting in unnatural outputs. For instance, in a Chinese cover letter writing task, ChatGPT generated responses with unnatural expressions, such as using "sincerely" in Chinese to sign off the letter. Fewer than 1% of the cover letters generated by ChatGPT included proper Chinese signoffs.

ChatGPT's unreliable performance is likely to be particularly evident in handling passages in dialects—a topic that remains largely underexplored in the field—presumably due to

limited training resources. Amid the paucity of research, Lilli (2023) found that ChatGPT-4's ability to handle different Italian dialects (Milanese, Sicilian, and Roman) varies depending on the assigned task. Among the tasks, ChatGPT-4 succeeded in comprehending dialects, translating them into standard Italian, and distinguishing between dialects. However, its accuracy markedly decreased when analyzing linguistic features or translating standard Italian into these dialects. It was concluded that ChatGPT-4 demonstrated stronger competence in the passive use of dialects (e.g., dialect recognition and comprehension) than in its active use (e.g., dialect production).

This finding suggests a limitation in ChatGPT's ability to generate texts that accurately reflect target dialects. However, given the current gap in knowledge, it is crucial to continue exploring this issue and advancing our understanding. This inquiry is particularly important for dialects of geopolitical significance, especially those for which experts are not readily available. Relying on speakers from different regions of the same language may not fully address the issue of limited resources due to inherent variations between dialects. For example, the two dialects of the Korean language—South and North Korean—exhibit significant differences in usage. *Language Magazine* (2015) reports that over 60% of vocabulary in professional contexts differs between the two dialects, while the difference ranges from 30% to 40% in everyday settings. This disparity presents a substantial obstacle for SK language experts in developing passages in the NK dialect, underscoring the need for ChatGPT's assistance in creating teaching and testing materials for dialects. Thus, the findings of this study provide valuable insight into the current state of ChatGPT's ability to recognize and produce non-English dialects, as well as its potential to support efforts in teaching and assessing them.

Research Questions

1. Does ChatGPT recognize NK passages when the level of passage difficulty is controlled for? If so, to what extent does it identify NK words and expressions, and provide their meanings accurately?
2. Does ChatGPT convert SK passages into NK passages when the level of passage difficulty is controlled for?

Methods

This study was conducted to support the development of NK passages for the Korean program at the U.S. Military Language Institute. All graduating students in the program are required to take a test containing NK passages and demonstrate at least Level 2 proficiency in Reading and Listening according to the Interagency Language Roundtable (ILR) scale. Learners at ILR Level 2 are expected to comprehend the straightforward spoken or written language samples on daily topics, including personal life, work, and current events (Interagency Language Roundtable, n.d.). To help students achieve this goal on their tests, it is necessary for instructors to develop new passages and continuously update existing resources. However, this poses a great challenge for the large body of Korean faculty—most of whom are from South Korea—due to the linguistic differences between the SK and NK dialects.

SK Dialect vs. NK Dialect

The linguistic divergence between SK and NK dialects has become evident in various areas since the separation of South and North Korea in 1945. One of the most notable differences would be vocabulary and its varied usage in different contexts (Berthelier, 2023). As shown in Table 12.1, the SK and NK dialects contain words that differ to varying degrees. For example, the SK and NK equivalents for the English words *Cheap* and *Vegetable* are more divergent than those for the word *Seashore* both orthographically and phonetically.

Table 12.1

Difference in Vocabulary between SK and NK

No.	English Word	SK		NK	
1	Cheap	싸다	[ssa-da]	눅다	[nuk-da]
2	Vegetable	채소	[chae-so]	남새	[nam-sae]
3	Seashore	바닷가	[ba-dat-ga]	바다가	[ba-da-ga]
4	Fish	생선 (for food)	[saeng-seon]	물고기	[mul-go-gi]
		물고기 (live fish in water)	[mul-go-gi]		

Note. The examples were selected from the Korean textbooks in the program.

Another important distinction in vocabulary between the dialects is that the same word can exist in both, yet it may convey different meanings depending on the context. For instance, in the SK dialect, the term for fish is divided into two types, fish for food and live fish in water, whereas no such distinction exists in the NK dialect. As a result, interpreting the term *물고기* [mul-go-gi] in NK passages requires an analysis of contextual cues to determine its specific meaning. These differences highlight the need for appropriate training and materials to support learners in accurately comprehending NK passages.

Materials

A total of 100 SK and 100 NK passages were selected from the program's Korean textbooks and adapted for the study. The passages were drawn from two semesters, consisting of four terms: Term 101 (21 passages), Term 102 (23 passages), Term 201 (36 passages), and Term 202 (20 passages). The difficulty of the passages increased progressively across the terms. The original SK passages were developed by SK teachers and subsequently rewritten in the NK dialect by NK teachers in the program, ensuring that the SK and NK passages contained identical information but differed only in dialect.

Although the same passages were used for both recognition and production tasks, the number of target words in each passage varied between the tasks. NK words and expressions were included in an average of 9.73% of each passage for the recognition task. However, the average percentage of target words requiring conversion into NK words in each SK passage increases to 11.40% in the production task (See Appendix A for details). This difference stems

primarily from the presence of English loanwords and Sino-Korean words—derived from Chinese characters—in SK passages. While these words are frequently used in SK, they are uncommon in the NK dialect. Consequently, NK teachers replaced them with words suitable for the NK dialect, resulting in their absence from NK passages in the recognition task. However, they remained in SK passages, increasing the total number of target words requiring translation into NK equivalents for the production task. For instance, the phrase *성공 스토리* (success story) in one of the SK passages includes the English loanword *스토리* [seu-to-ri], which is acceptable in SK, whereas, in the production task, the word must be replaced with the native Korean word, *이야기* [iyagi].

Procedures

A subscription version of ChatGPT-4o was used to perform the recognition and production tasks, focusing on vocabulary and expressions necessary to complete them (see Appendices B and C for the prompts). ChatGPT-4o was initially tasked with distinguishing between SK and NK written passages, then identifying NK words and expressions when the passages were recognized as NK dialect, and finally providing their meanings. After completing the recognition task, it was asked to convert SK passages into NK passages without altering the content. Passages with different dialects and difficulty levels were mixed and randomly presented to ChatGPT-4o for both tasks. It performed the recognition task once, but in the production task, a second attempt was made in eight instances where content changes were observed. In these cases, the researcher prompted ChatGPT-4o to perform the task again, leading to immediate error correction on the second attempt.

Data Analysis

ChatGPT-4o's responses to the recognition and production tasks constitute the data for this study. To address the first research question, the following analyses were conducted. First, the cases in which ChatGPT-4o correctly recognized each dialect were counted. Next, a multiple logistic regression model was conducted with two independent variables, dialect and passage difficulty, and the dependent variable, ChatGPT-4o's correct or incorrect judgment of the given dialect. This regression model, suitable for categorical dependent variables, was selected to examine the unique contribution of dialect to the variance of the dependent variable while controlling for the effect of passage difficulty (Larson-Hall, 2015). Additionally, NK word identification and meaning accuracy rates were calculated by comparing the NK words identified by ChatGPT- 4o with those converted by NK teachers. Finally, errors observed in the NK word identification tasks were analyzed to reveal noticeable patterns.

For the second research question, the conversion accuracy rate was calculated based on comparisons between ChatGPT-4o's and NK teachers' converted words. Next, a linear regression analysis, designed for analyzing data with a continuous dependent variable, was conducted with passage difficulty as the independent variable and the conversion accuracy rate as the dependent variable. An analysis of conversion errors was conducted to identify patterns.

Results

The first research question asks whether ChatGPT-4o can recognize NK passages when the level of passage difficulty is controlled for and, if so, to what extent it can identify NK words and expressions and provide their meanings. Table 12.2 shows that ChatGPT-4o correctly recognized 98 SK passages (98%) and 97 NK passages (98%) out of 100, demonstrating its excellent performance in distinguishing between the dialects across varying levels of passage difficulty, ranging from 95% to 100%. The multiple logistic regression results also found no significant difference between dialects and across different levels of passage difficulty, with all *p* values greater than .05 (Table 12.3).

Table 12.2

Dialect Recognition Accuracy Rate by Term

	Term 101		Term 102		Term 201		Term 202		Total	
	SK	NK	SK	NK	SK	NK	SK	NK	SK	NK
Count	21	20	23	23	36	35	18	19	98	97
%	100	95.24	100	100	100	97.22	90	95	98	97

Note. The total number of passages selected from each term: Term 101 = 21 passages, Term 102 = 23 passages, Term 201 = 36 passages, and Term 202 = 20 passages

Table 12.3

Multiple Logistic Regression Model Results for Passage Difficulty Levels and Dialects

	β	SE	Z	*p*
Intercept	7.57	651.72	0.012	0.991
Term 102	-4.45	582.92	-0.008	0.994
Term 201	-9.30	1303.44	-0.007	0.994
Term 202	10.66	1748.76	0.006	0.995
Dialect	0.42	0.93	0.45	0.648

Note. Reference Level: Term 101

Along with its strong performance in recognizing dialects, ChatGPT-4o was generally proficient in identifying NK words and expressions (Table 12.4). On average, 87.77% of NK words were correctly identified in each passage, with a range from 84.61% to 91.08%. Its performance was fairly consistent across different levels of passage difficulty.

Table 12.4

Average NK Word/Expression Identification Rate

Term	NK Word/ Expression in Each Passage		Correctly Identified NK Words/Expressions	
	Average Count	%	Average Count	%
101	5.14	9.64	4.33	87.79
102	5.91	8.99	5.30	89.81
201	5.42	10.22	4.56	84.61
202	3.95	9.89	3.60	91.08
Total	5.11	9.66	4.49	87.77

An analysis of the meanings provided by ChatGPT-4o also found that it was highly accurate in identifying correct meanings (Table 12.5). ChatGPT-4o offered correct meanings, including either English meanings or SK equivalents, in 95.36% of cases. Most often, both English meanings and SK equivalents were given (77.63%), with instances of English meanings alone (6.94%) and SK equivalents alone (10.74%). In a few cases, it only addressed the NK writing style of the passage without providing any meanings (3.80%). Inaccurate meanings were rarely observed in the study (0.89%).

Table 12.5

NK Word Meanings Provided by ChatGPT

	Meaning Alone		SK Equivalent Alone		Both		Style Alone		Inaccurate Meaning	
	Count	%	Count	%	Count	%	Count	%	Count	%
101	3	3.30	9	9.89	72	79.12	7	7.69	0	0.00
102	9	7.38	14	11.48	97	79.51	2	1.64	0	0.00
201	17	10.43	20	12.27	123	75.46	1	0.61	2	1.23
202	2	2.82	5	7.04	55	77.46	7	9.86	2	2.82
Total	31	6.94	48	10.74	347	77.63	17	3.80	4	0.89

Note. The total number of NK words/ expressions = 447 (Term 101=91; Term 102=122; Term 201=163; Term 202 = 71)

Although ChatGPT-4o successfully identified most NK words and expressions, any errors were still analyzed to identify potential patterns. Table 12.6 shows that more than half of the errors occurred when it failed to correctly examine the context where the target words were used (58.06%), whereas the remaining errors were limited to the cases of a mixture of correct or incorrect identification (19.35%) or those where it simply missed the NK words (22.22%).

Table 12.6

Types of NK Word Identification Errors

	Inconsistent Identification	Missed	Inappropriate Context	Total
Count	12 (19.35%)	14 (22.22%)	36 (58.06%)	62

An example of ChatGPT-4o misjudging contextual cues surrounding a target word is observed in the following sentence:

(1) [...] 중심선을 넘어 마주오던 뻐스를 […]

(English Translation: the bus coming toward us, crossing the center line)

The underlined Korean word, 중심선 [jungsim-seon], translates to "the center line" in English. In the NK passage reporting a car accident, 중심선 is used to describe the road's center line. However, in the SK dialect, 중심선 is not typically used for the center line on the road; instead, 중앙선 [jungang-seon] is considered more appropriate. While 중심선 exists in SK, it is more commonly used to describe a circle's center line, typically observed in math class. ChatGPT-4o should have recognized the inappropriate use of 중심선 in this context within the SK dialect but failed to identify it as a word from the NK dialect.

However, it should be noted that ChatGPT-4o does not always misjudge the words existing in both SK and NK dialects. One of the examples that elicited its accurate evaluation of the context was 물고기 [mul-go-gi] (fish). Consider the use of 물고기 in the following sentence:

(2) [...] 물고기를 자주 먹습니다. (English Translation: [They] often eat fish.)

This sentence came from one of the NK passages that describe people's diets. In the SK dialect, the use of 물고기 in this context would sound unnatural since 물고기 typically refers to live fish in water, not fish for food. ChatGPT-4o, in contrast to the previous example, evaluated the surrounding context of the target word correctly, included it in the list of NK words, and suggested an SK alternative as shown in the following response (3):

(3) ChatGPT-4o Response: 물고기 - While "물고기" (fish) is used in both Koreas, in South Korea, simply "생선" would often be used in the context of food.

These examples show that ChatGPT-4o has some competence in evaluating the context, yet its ability still seems to be limited, failing to yield consistent results.

The second research question addresses ChatGPT-4o's ability to convert the SK dialect to the NK dialect when the passage difficulty is controlled for. Overall, ChatGPT-4o's competence declined notably in converting the SK dialect to the NK dialect, especially when compared to its strong performance in dialect recognition. The accuracy rates remained low, with an average rate of 31.39% (Table 12.7).

Table 12.7

Accuracy Rate of Conversion from SK Dialect to NK Dialect

Term	Average Number of Target Words in Each Passage	Count	%
101	6.19	2.00	32.60
102	7.09	2.00	27.72
201	6.28	1.94	30.96
202	4.40	1.70	35.83
Total	6.07	1.91	31.39

This decline was observed consistently across the terms, with the highest accuracy rate of 35.83% in Term 202, well below 50% (Table 12.7). The consistently low performance aligns with the findings from the linear regression models, which revealed no significant difference in ChatGPT-4o's conversion accuracy across the terms as indicated by their *p* values remaining above .05 (Table 12.8).

Table 12.8

Linear Regression Model Results for Passage Difficulty Levels

	β	SE	Z	*p*
Intercept	87.35	1.81	48.184	< .001
Term 102	-0.24	3.84	-0.063	0.950
Term 201	2.49	3.62	0.689	0.493
Term 202	4.35	3.39	1.285	0.202

Note. Reference Level = Term 101

The cases of ChatGPT-4o's conversion failures were further analyzed to identify potential patterns. The analysis showed that in most cases, ChatGPT-4o did not attempt to make any changes to the target words, making up 71.81% of conversion errors (Table 12.9). This result reveals ChatGPT-4o's limitation in recognizing the SK words used differently from NK words, resulting in no attempt to make changes. This limitation contrasts with its strong ability to identify NK words and their meanings in NK passages, demonstrating that its proficiency is restricted to a single direction: NK to SK, rather than the reverse.

Table 12.9

Types of Conversion Failures

Term	No Change		Inaccurate		Mixed		Total
	Count	%	Count	%	Count	%	
101	61	69.32	25	28.41	2	2.27	88
102	87	74.36	30	25.64	0	0.00	117
201	115	73.72	41	26.28	0	0.00	156
202	35	64.81	19	35.19	0	0.00	54
Total	298	71.81	115	27.71	2	0.48	415

Note. Mixed = A mixture of accurate and inaccurate production of the target words/expressions

Inaccurate NK equivalents provided by ChatGPT-4o accounted for 27.71% of the cases, while mixed cases, where the same target words were converted inconsistently into correct equivalents across passages, were rarely observed (0.48%).

The incorrectly converted cases exhibit various types of errors. One type involves minor inaccuracies in converted words. For example, the SK word 납치 (kidnapping) was changed to 렵치, which is close to the target word 랍치, differing by only a single vowel, ㅏ versus ㅕ. Another error occurred when ChatGPT-4o's attempt at a literal translation was inappropriate for the context. The target word 사격교관 (a shooting military instructor) was originally converted by NK teachers to 사격 지휘관 (a shooting military commander) to fit the context, whereas ChatGPT-4o translated "an instructor" literally into the NK term for a teacher, *교원*, resulting in *사격교원* (a shooting teacher).

A further error involved ChatGPT-4o using a more generic term instead of the specific target word. For instance, the SK word *공무원* (a government official) was changed to *일군*—likely intended to be *일꾼* by ChatGPT-4o—which means a worker, a more general term, although the specific NK equivalent *정무원* exists in the dialect. Similarly, the English loan word *테니스* (tennis) in the SK passage was converted to *뿔운동*, which means *ball game*, instead of the specific term *정구* in the NK dialect. These examples demonstrate that while ChatGPT can recognize some South Korean words not used in the North Korean dialect, its ability is too limited to consistently produce correct North Korean equivalents, emphasizing the need for further training.

Discussion

Overall, ChatGPT-4o was highly capable of recognizing the NK dialect, identifying NK words in each passage, and providing their meanings or SK equivalents; however, its performance declined notably when tasked with converting the SK dialect to the NK dialect. This gap aligns with Lilli's (2023) findings that ChatGPT 4 outperformed in tasks involving dialect recognition and comprehension compared to tasks requiring active use of target Italian dialects, such as translating standard Italian into each dialect.

The consistent findings of ChatGPT 4's strong performance in recognizing and comprehending foreign dialects demonstrate the rapid progress in AI technology, particularly in its ability to process language input—even for low-resource languages like NK dialect. Berthelier (2023), while noting the dearth of NK language resources available online, raised the possibility of developing an NLP model capable of processing both SK and NK dialects primarily using SK language resources—a high-resource language—with only a small supplement of NK resources published by the SK government and institutes. In line with these expectations, ChatGPT-4o demonstrated high proficiency in identifying NK words and providing their meanings, though the specifics of its training resources remain unclear.

Despite significant progress in recognizing the dialect, ChatGPT-4o's ability to accurately evaluate the context surrounding target words remains in question. In the study, its performance was inconsistent in identifying North Korean words shared with the South Korean dialect but used in different contexts. This task requires ChatGPT-4o to consider the context in which the target words are used, potentially overreaching its capabilities. Berthelier (2023) addressed the challenges posed by differing contextual embeddings of the same words between the SK and NK dialects, adding complexity to developing an NLP model. Given that analyzing contextual cues to derive accurate meaning is essential for fully understanding a language, this challenge will likely extend to other languages and dialects. As such, it is an important task to address, warranting continued research and efforts to train AI.

Another limitation of ChatGPT-4o's competence, consistently observed in the current and Lilli's (2023) studies, is its ability to produce the target dialect. In both studies, ChatGPT 4 struggled to translate standard Italian or the SK dialect, higher-resource languages, into their target dialects, which are more limited in language resources. This low performance contrasts with ChatGPT's strong ability to generate English texts (Aryadoust et al., 2024; Shin & Lee, 2023; Xiao et al., 2023), presenting a clear gap between its ability to generate texts in English and in non-English languages, especially dialects, and suggesting that great caution be taken when leveraging ChatGPT to develop passages in dialects. Overall, these findings demonstrate that ChatGPT-4o has made significant progress in comprehending foreign languages, including dialects classified as low-resource languages; however, this advancement has not directly translated into improved production skills.

Conclusion

The findings, revealing ChatGPT-4o's capabilities and limitations, offer valuable insights into its potential applications in language teaching and testing. ChatGPT-4o can be useful for verifying NK passages found online or creating vocabulary lists based on NK texts. However, caution is strongly advised when using ChatGPT-4o to develop NK passages, especially when translating SK texts into the NK dialect. This process requires additional verification through resources such as published research and materials on the NK dialect to ensure accuracy.

Diagnosing the current state of ChatGPT-4o's contributions to language teaching and testing presents the next steps and identifies areas requiring further research. First, ChatGPT-4o's ability to comprehend contextual meaning should be explored through more empirical data. While the current study included instances where ChatGPT-4o needed to analyze contextual cues

to identify NK words, a more precise evaluation of its contextual comprehension requires research with a larger dataset that focuses on contextual meanings.

Additionally, continuous efforts are needed to optimize the use of ChatGPT-4o in developing passages in dialects. Although the current study revealed its limitations in translating between dialects, alternative approaches to enhance its performance should be explored, such as developing NK passages based on a given list of NK words. Investigating various methods and fine-tuning prompts could potentially improve its accuracy in production. This endeavor should extend beyond ChatGPT-4o, involving active research into the capabilities of other AI tools, such as Gemini and Clova X, to further support language teachers and test developers, ultimately providing high-quality language input for learners.

This study calls attention to the importance of integrating AI technology into the development of teaching and testing materials for low-resource languages—especially non-English dialects—to address current resource limitations and identify areas for improvement, ultimately making AI tools such as ChatGPT even more effective for language teaching and assessment.

References

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, *6*, 100204. https://doi.org/10.1016/j.caeai.2024.100204

Bachman, Lyle F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Berthelier, B. (2023). Division and the digital language divide: A critical perspective on natural language processing resources for the South and North Korean languages. *Korean Studies*, *47*(1), 243-273. https://doi.org/10.1353/ks.2023.a908624

Interagency Language Roundtable. (n.d.). *ILR skill level descriptions.* https://perma.cc/BM8M-8NJP

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537-550. https://doi.org/10.1177/00336882231162868

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). *ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning*. ArXiv. https://doi.org/10.48550/arXiv.2304.05613

Language Magazine. (2015, March 30). *Bridging the North and South Korean language divide*. https://perma.cc/8SJ4-F44P

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Lilli, S. (2023). ChatGPT-4 and Italian dialects: Assessing linguistic competence. *Umanistica Digitale*, *7*(16), 235-263. https://doi.org/10.6092/issn.2532-8816/18221

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://openai.com

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology, 27*(3), 27–40. https://doi.org/10.64152/10125/73530

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 610–625). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.52

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023). *Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs*. ArXiv. https://doi.org/10.48550/arXiv.2305.16339

Appendix A

Table A1

NK Word/ Expression Distribution for Recognition Task

Term	Word Count	NK Words/ Expressions		Repetition of Target Words/ Expressions	
		Count	%	Count	%
101	55.33	5.14	9.56	1.57	28.64
102	69.17	5.91	8.99	1.70	27.13
201	55.31	5.42	10.22	2.19	39.72
202	42.30	3.95	9.89	1.25	28.67
Total	55.90	5.11	9.66	1.76	32.29

Note. Word Count = The average word count in each passage

Table A2

Target Word/ Expression Distribution for Production Task

Term	Word Count	Target Words/ Expressions for Conversion		Repetition of Target Words/ Expressions	
		Count	%	Count	%
101	55.33	6.19	11.74	1.52	23.82
102	69.17	7.09	10.72	1.70	23.83
201	55.31	6.28	11.77	2.00	31.38
202	42.30	4.40	11.14	1.15	22.55
Total	55.90	6.07	11.40	1.66	26.29

Note. Target Word/ Expressions = SK words/ expressions not used in NK

Appendix B

Prompt for Recognition Task

I will give you a passage. Tell me whether it is written in a South Korean or North Korean style. Evaluate it based only on the use of words or expressions in the passage. Do not consider sentence length or word spacing. If it is written in a North Korean style, please provide a list of North Korean words and expressions along with their meanings.

Appendix C

Prompt for Production Task

I will give you a South Korean passage. [The passage is a casual speech by a female student.] Could you rewrite it in a North Korean style using North Korean words and expressions? Keep the original information the same as in the South Korean passage. Do not add or change the original meaning. Do not paraphrase the sentences. Replace the South Korean words or expressions with North Korean ones only when necessary. [Maintain the casual speech and the voice of the writer, who is a female.] Additionally, provide a list of changes you made.
Note: The content within the brackets varies depending on the passage.

About the Author

Jean Young Chun holds a doctoral degree in Second Language Studies and currently serves as an assessment specialist in the Curriculum Development (CD) Division at the Defense Language Institute Foreign Language Center (DLIFLC) in Monterey, CA. Her previous career focused on language teaching and testing at secondary and post-secondary institutions. Her research interests include L2 assessment and pragmatics.

Disclaimer

- This chapter has been approved for public release by the Defense Language Institute Foreign Language Center's Public Affairs Office. For verification please e-mail: mpao@dliflc.edu
- Contents of this chapter are not necessarily the official views of the Defense Language Institute Foreign Language Center, nor are they endorsed by the Department of the Army, the Department of Defense, or the U.S. Government.
- All third party products/materials featured in the presentation remain the intellectual property of their respective authors/originators. Use of outside materials is done under the fair use copyright principle, for educational purposes only.
- The content of this chapter is the sole responsibility of the author(s).

Chapter 13

Comparing Human and Generative AI Analysis of Emotions in U.S. Abortion Tweets

Şebnem Kurt and Mahdi Duris
Iowa State University

This study investigates the use of ChatGPT-4o to analyze contributors' emotions and stances regarding abortion as expressed in U.S. abortion-related tweets. The study compared two human analysts' and ChatGPT's coding of a dataset of 100 tweets from Boston, Massachusetts, collected between September 2022 and January 2023. Codes for affect were defined based on Martin and White's (2005) Appraisal Framework, focusing on the AFFECT subcategory to identify positive, negative, and neutral emotions, and stance codes were defined using Du Bois' (2007) stance-taking analysis to classify pro-abortion, anti-abortion, or neutral stances. The motivation for this research stems from the challenges inherent in human coding of expressions of subjectivity including its time-consuming nature, low interrater reliability, and potential biases arising from factors like insufficient coder training, varying familiarity with the cultural context, and effects stemming from differences in coders' gender identities. By exploring the efficacy of ChatGPT in analyzing emotional content, this study aimed to assess the potential of GenAI tools to support discourse analysis of social media. Findings reveal only fair agreement in emotion classification (Kappa=0.355 between Human 1 and Human 2, Kappa=0.283 between Human 1 and ChatGPT, and Kappa=0.322 between Human 2 and ChatGPT) and in stance classification (Kappa=0.402 between Human 1 and Human 2, Kappa=0.283 between Human 1 and ChatGPT and Kappa=0.468 between Human 2 and ChatGPT). Overall reliability across coders, as measured by Krippendorff's Alpha, was also low (Alpha=0.319), further underscoring the challenges of achieving high consistency in subjective coding tasks such as emotion and stance analysis. Chi-square

Cite as: Kurt, S., & Duris, M. (2025). Comparing human and Generative AI analysis of emotions in U.S. abortion tweets. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 252-273). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.13
©2025 Kurt and Duris. Published under a CC BY license.

analysis further highlights a significant difference in how ChatGPT and human coders categorized emotions (positive, negative, and neutral) in abortion tweets. These results provide valuable insights into the capabilities and limitations of ChatGPT-4o in analyzing subjectivity in social media content, particularly within the polarizing content of abortion discourse.

Introduction

Human-Computer Interaction (HCI) has undergone significant transformations since Licklider (1960) first imagined a symbiotic relationship between humans and computers. The introduction of early graphical user interfaces, the development of the mouse (Myers, 1998), and the widespread adoption of the internet (Hartson, 1998) marked the initial steps towards actualizing this vision. More recently, innovations such as voice assistants like Siri and Alexa (Hoy, 2018) and advancements in conversational AI (Casheekar et al., 2024) have further expanded these possibilities and prompted discussions about whether machine intelligence will ever be able to interpret commands and grasp the subjectivity inherent in human communication, including emotions and stances.

While technological progress in artificial intelligence has recently advanced, critical concerns linger regarding its ability to accurately identify emotional and ideological positions in text, particularly with controversial societal topics. Sentiment analysis has made significant progress in automated text classification (Taufek et al., 2021), yet numerous studies emphasize AI's shortcomings in performing "critical and contextualized analyses" (Curry et al., 2024, p. 2). Research that introduces qualitative feedback loops, where AI outputs are refined through iterative prompting and inductive coding (Hämäläinen et al., 2023; Rahman et al., 2023), highlights the critical role of human oversight. This oversight is especially crucial when researchers are investigating culturally nuanced or subtly expressed sentiments because a purely automated approach falls short of capturing the complexities of human communication. Despite the need for human oversight of AI emotion and stance analysis, few studies have included a qualitative dimension to the analysis. Additionally, there is a need for datasets that reflect contextual variability, which plays a key role in both human and AI accuracy.

Social media platforms like Twitter, rich with opinionated content, offer an opportunity to explore how AI coding of subjective expressions aligns with human evaluations. This study examines how AI evaluates human emotions and stances in text, focusing on U.S. abortion-related tweets using a mixed-methods approach.

Literature Review

A considerable amount of scholarly research on social media discourse has concentrated on the microblogging platform formally known as Twitter, which serves as a valuable repository for immediate and often dynamic exchanges. Many studies show that data from Twitter can be used effectively to develop algorithms for stance detection and sentiment analysis, thanks to the platform's concise and public format. For instance, Mohammad et al. (2016) used a detailed, manually annotated Twitter dataset to improve computational stance-detection tools. Their study demonstrated that Twitter data can effectively provide researchers with opinion-rich content.

From a linguistic perspective, Wikström (2019) highlighted Twitter's capacity for providing animating reported speech and shaping social identity, while Poletto et al. (2017) discussed how Twitter communications can help uncover the characteristics of hate speech directed at minority groups. Additionally, Giaxoglou and Spilioti (2020) examined users' engagement with brief narratives and visual stories. Their analysis highlighted the interplay between social and cultural contexts in stance-taking on the microblogging platform.

As AI-based methods for sentiment and stance detection continue to evolve, researchers have engaged in discussions about whether computational tools can effectively interpret both explicit and implicit meanings on equal grounds. On the one hand, large language models have shown relatively high accuracy when stances are clearly articulated, as demonstrated by Gilardi et al. (2023) and Zhang et al. (2023). However, implicit attitudes, those conveyed through context, inference, or subtle cues, present a greater challenge. Li and Conrad (2024) noted that even human annotators struggle with implicit stance detection. These parallels between machine and human challenges reveal the importance of context and suggest that the "humanity" of the dataset, whether rich in explicit statements or heavily reliant on implied meanings, can significantly affect the quality of annotations.

Research examining online debates on polarizing issues such as abortion or gun rights underscores the importance of both implicit and explicit stance-taking in shaping public discourse. For instance, Burnap et al. (2015) used computational methods to analyze social tension on Twitter, illustrating how linguistically charged topics are framed in online interactions. Similarly, Sharma et al. (2017) conducted a Critical Discourse Analysis of abortion-related content and demonstrated how specific discursive strategies influence public attitudes and moral positioning on this contentious issue. More recently, studies have begun comparing the effectiveness of AI-driven tools and human coders in identifying nuanced emotional and ideological cues in social media content. These studies point to the value of mixed-methods approaches that integrate computational techniques with qualitative insights to better understand how both humans and AI interpret emotionally charged and polarized discourse (Ballesteros et al., 2024; Elyoseph et al., 2024; Huang et al., 2023; Machová et al., 2023). As large language models such as the one powering ChatGPT become more widely used in these contexts, new questions arise about their reliability and interpretive capabilities.

Although recent research has explored the potential of AI in stance detection, less is known about the performance of large language models (LLMs) like ChatGPT in this area. In particular, few studies have examined how LLMs handle both emotional expression and stance within highly polarized sociopolitical contexts. Therefore, there is a gap in understanding whether AI-generated annotations align sufficiently with human coders to support their use in applied linguistics and discourse analysis, especially in the context of contentious topics such as abortion. To fill this gap, we selected Martin and White's (2005) Appraisal framework and Du Bois' (2007) stance-taking framework to analyze Twitter data using both human and AI coders. The Appraisal framework allows for a fine-grained analysis of emotional expression by systematically identifying and categorizing affective meanings at the word and sentence level in tweets. It provides a detailed methodology for dissecting complex emotional expressions in text,

guiding researchers to identify specific lexical markers (i.e., distinguishing negative words such as [“devastated” and ”angry”] from positive words such as [“hopeful” and ”relieved”]), which is essential for analyzing the nuanced and often subtle affective markers found in abortion-related tweets. Similarly, Du Bois’ stance-taking framework enables a systematic examination of how users position themselves in discourse through alignment, evaluation, and positioning moves. By conceptualizing stance as a triangular relationship between the speaker, the issue (e.g., abortion), and other participants, the framework supports the classification of posts as pro-abortion, anti-abortion, or neutral. This, in turn, reveals how individuals express their evaluations and align, or fail to align, with opposing perspectives.

Frameworks such as Critical Discourse Analysis (CDA), rhetorical argument analysis, and domain-specific computational methods can shed light on how emotion and stance coexist in abortion-related tweets, but Martin and White’s Appraisal framework and Du Bois’ stance-taking framework offer a more fine-grained approach. These two frameworks not only help analyze how emotional content is expressed but also how individuals position themselves in relation to polarizing issues. They allow us to distinguish, for example, whether a speaker’s negativity stems from moral condemnation (i.e., negative JUDGEMENT) or personal distress (i.e., negative AFFECT), and to trace how individuals align, or misalign, with others in ongoing Twitter conversations. Such detailed, layered insights are often overlooked by broader computational or ideology-driven methods.

By applying these frameworks, this study offers new insights into how generative AI models like ChatGPT interpret AFFECT and stance in polarizing social media discourse, and how their performance compares to that of trained human coders. Building on prior research that highlights AI’s promise and limitations in emotion and stance detection, our study offers insights into whether ChatGPT, a specific application of LLMs, can code subjective expressions of emotion and stance with sufficient agreement to be useful in discourse analysis within applied linguistics, particularly in the language of contentious public debates. The research was guided by the following research questions:

1. How frequent are positive, negative, and neutral emotions expressed in U.S. abortion-related tweets from Boston?
2. How frequent are pro-abortion, anti-abortion, or neutral stances displayed in these tweets?
3. To what extent does ChatGPT agree with human coders in identifying emotions and stance in abortion-related tweets, and how does this agreement compare to inter-rater agreement between human coders?
4. Are there statistically significant differences between human coders and ChatGPT in coding for emotions and for stance within these tweets?

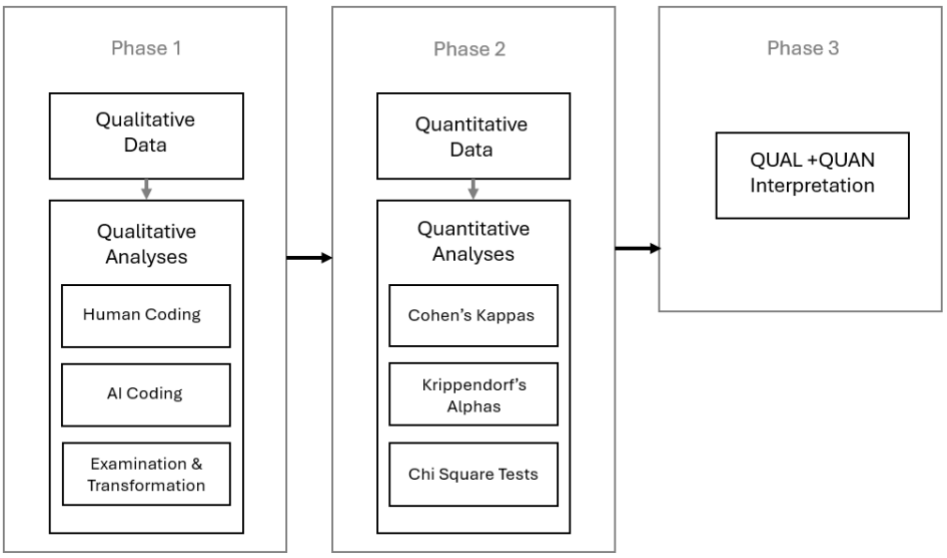
Methods

We employed a three-phase mixed-methods design (Creswell & Plano Clark, 2017) with qualitative data and analyses in Phase 1, quantitative data and analyses in Phase 2, and interpretation of the analyses from the two phases in Phase 3, as shown in Figure 13.1. The unit

of analysis throughout was the body of each of 100 tweets, consisting of the linguistic comment posted by users engaging with the topic of abortion on Twitter. A tweet can consist of a single word or a series of statements, but the complete unit, regardless of length or syntactic form, received only one code from each of two human coders and from ChatGPT-4o in Phase 1. The qualitative analysis of the data was carried out by two human coders and by ChatGPT-4o, who each assigned an AFFECT code and a stance-taking code to each tweet, guided by Martin & White's (2005) Appraisal and Du Bois' (2007) Stance-Taking frameworks, respectively. The three sets of codes for emotions and the three sets for stance were nominal data that were transformed into quantitative data for statistical analysis in Phase 2 of the study. In Phase 2, Cohen's Kappa was used to calculate agreement between each pair of coders (three pairs of AFFECT codes and three pairs of stance codes) while correcting for agreement occurring by chance. Krippendorff's Alpha was used to obtain a single agreement statistic across the three coders to assess overall reliability for coding AFFECT and for coding stance. Additionally, chi-square tests were conducted to test for statistically significant differences in the assignment of codes between humans and ChatGPT on the tweets that the humans had agreed upon. This helped evaluate not only consistency but also details about coding similarities and differences in coding by the human coders and ChatGPT. In Phase 3 of the study, interpretations are drawn about the use of ChatGPT-4o for assistance in qualitative coding in applied linguistics based on statistical findings about agreement obtained in the use of the qualitative frameworks for AFFECT and stance.

Figure 13.1

Mixed-Methods Research Design



Data

This study analyzes abortion-related tweets collected from geotagged locations in the 50 U.S. state capitals. Using a Python program and Tweepy, a Python library that enables access to the Twitter Application Programming Interface (API), we scraped tweets containing the keyword

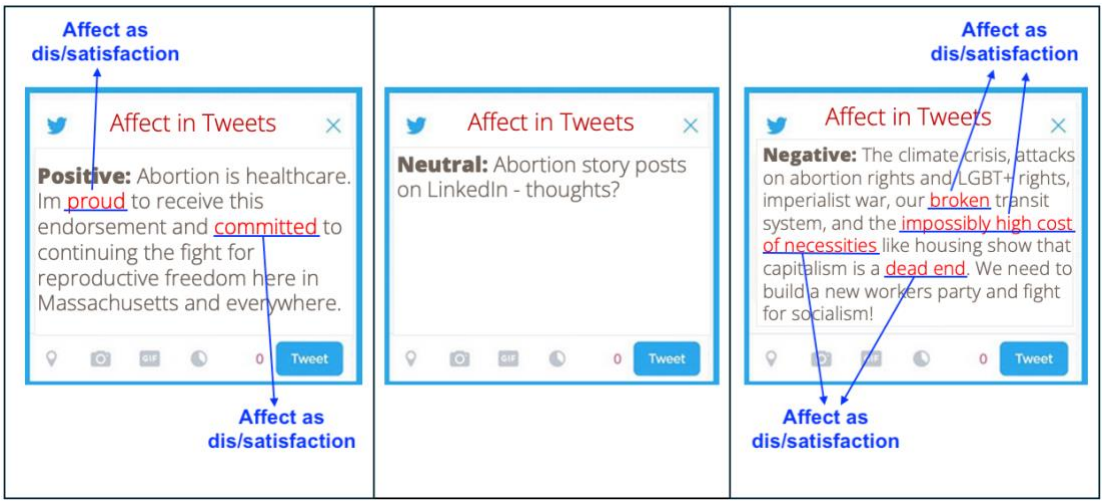
"abortion" between September 2022 and January 2023. The initial dataset comprised over 100,000 tweets. After cleaning the data using natural language processing (NLP) techniques, we retained the tweet text and date for each entry. The cleaned data was saved as separate CSV files for each city. Given the labor-intensive nature of manual analysis, we selected a manageable subset for in-depth examination. For this study, we randomly selected the Boston dataset, consisting of 2,565 tweets, for further analysis. Tweets were coded using the Appraisal framework for emotion (positive, negative, neutral) and Stance for position (pro-abortion, anti-abortion, neutral). These categorical codes were then converted into numerical values to conduct statistical analyses, including interrater reliability (Cohen's kappa and Krippendorff's alpha) and chi-square tests.

Coding Emotion

The Appraisal framework (Martin & White, 2005) is a systemic functional linguistic approach that categorizes evaluative language according to how it conveys emotions, opinions about human behavior, and evaluations of personal feelings and phenomena. For our emotion analysis, we focused on AFFECT, under the subsystem of Attitude, which captures positive and negative linguistic expressions of emotions within three categories, each with both a positive and a negative valence: un/happiness, in/security, and dis/satisfaction (Martin & White, 2005). Each of these categories of AFFECT can be expressed by a variety of different linguistic devices, which have to be interpreted in terms of their meaning in the context where they are used so there is not a complete list of indicators for positive and negative AFFECT that coders can apply formulaically. Figure 13.2 shows the AFFECT analysis coding by one human on three tweets from the dataset. The coder identified the tweet on the left by the words "proud" and "committed," which they judged to be positive words falling within the AFFECT category of dis/satisfaction. The tweet on the right was judged to indicate negative AFFECT in the same category, using words and phrases such as "broken," "impossibly high cost of necessities," and "dead end." The tweet in the middle lacks language indicating positive or negative emotions and was therefore coded as neutral. Since Martin and White did not provide an exhaustive list of linguistic indicators for AFFECT, we used their framework to build a conceptual guide to establish a shared understanding of AFFECT for the human and ChatGPT coders. For the purposes of this study, however, tweets were only coded as positive, negative, and neutral rather than being assigned to specific AFFECT categories.

Figure 13.2

Example of How Appraisal Framework was Applied to the Abortion Tweets

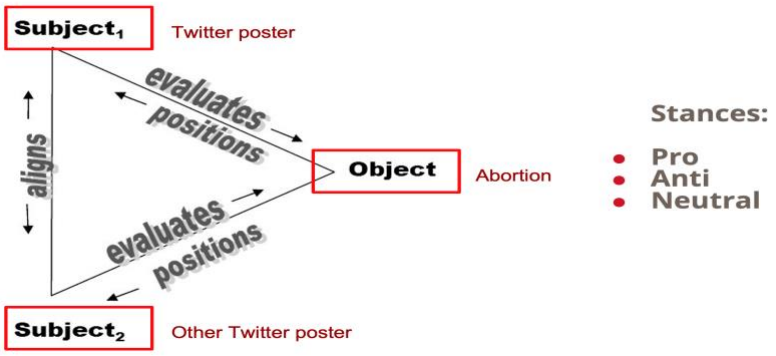


Coding Stance-Taking

Du Bois (2007) defines stance-taking as the use of language to evaluate objects, position subjects, and align with others. Stance-takers can express evaluations through evaluative language (e.g., adjectives like "proud" or predicates like "committed") that signals their positioning and alignment with other subjects (e.g., toward groups like "women" or concepts like "reproductive freedom"). Positioning is accomplished through language that situates social actors as subjects (e.g., pronouns such as "I" or nominal groups such as "people" or "our country") in relation to predicates expressing emotional or epistemic stances toward stance objects, or the words and phrases that language users choose to evaluate (e.g., "abortion law," "rights," or "healthcare"). Alignment occurs when language users adjust their relationship to others' stances by agreeing or disagreeing. Together, evaluations, positionings, and alignments create a triangular relationship, reflecting the stance-taker's identity and perspective (Du Bois, 2007). Figure 13.3 illustrates the application of stance-taking in this study.

Figure 13.3

Application of Du Bois' Stance-Taking Framework to Analyze Twitter Posts on Abortion.



In Figure 13.3, Subject 1 is the Twitter poster, Subject 2 is another poster, and the Object is the topic of abortion. We analyzed how each poster evaluated the object (abortion), positioned themselves, and aligned or misaligned with other Twitter posters. The identified stances were pro-abortion, anti-abortion, and neutral. The example in Figure 13.4 shows examples of how three tweets were coded for stance by one human coder.

Figure 13.4

Example of How Stance Framework was Applied to the Abortion Tweets.



In Figure 13.4, the tweet on the left is coded as pro-abortion because it expresses a stance of evaluation by equating abortion with "healthcare" and describing it as "reproductive freedom," thus positively evaluating the stance object (abortion). The author positions themselves as the stancetaker through the use of the pronoun "I," and adopts a committed stance toward the stance object, the fight for abortion rights, by using the predicates "proud" and "committed." These lexical choices, which are underlined in the tweet body in Figure 13.4, signal both an evaluative alignment and a clear positioning within the discourse of reproductive freedom. This positioning also shows alignment with others who advocate for reproductive freedom. Similarly, words expressing a negative stance on abortion are underlined in the anti-abortion tweet on the right, which evaluates abortion as harmful or wrong, again demonstrating evaluation and positioning while aligning with anti-abortion perspectives. In contrast, the neutral tweet in the middle lacks explicit stance markers; there are no evaluative expressions or clear indications of positioning or alignment, leading to its classification as neutral within the stance triangle framework. For the purposes of this study, each tweet was coded as pro-abortion, anti-abortion, or neutral, based on the stance it expressed toward abortion.

Procedures

In Phase 1 of the research, the researchers gathered, compiled, and cleaned the tweets, and a sample of 100 of them was coded first by the two researchers and then by ChatGPT. The human coders were both applied linguists with extensive training in discourse analysis, having completed PhD-level courses where they engaged with the theoretical foundations of appraisal analysis and stance-taking frameworks and had applied them in various research contexts. However, the coders' familiarity with the US political context, as well as their own political views, may have influenced their interpretation of appraisal and stance-taking strategies.

ChatGPT-4o coded the same 100 tweets using both the AFFECT and stance-taking frameworks. Since ChatGPT-4o is a pre-trained language model and not specifically fine-tuned for this task, we employed a structured prompting approach to guide its coding process. First, we introduced the theoretical background of the Appraisal framework by providing definitions and explanations of its three domains: ATTITUDE, ENGAGEMENT, and GRADUATION. We then instructed ChatGPT to focus exclusively on the AFFECT subsystem under the ATTITUDE system to establish the scope of its coding. To check comprehension, we asked it to summarize the key aspects of the framework. Finally, we presented tweets one by one and prompted ChatGPT-4o to classify each as expressing positive, negative, or neutral AFFECT and required justifications based on the provided theoretical background. ChatGPT-4o also coded the same 100 tweets using the stance-taking framework outlined by Du Bois (2007). A similar structured prompting approach was used to adapt the framework for the model. We introduced the stance triangle, including its three components: evaluation of a stance object, positioning of the speaker or writer, and alignment with others. We provided ChatGPT with definitions and examples of how stance can be expressed through linguistic features such as adjectives, predicates, and pronouns. To check comprehension, we asked it to summarize the key aspects of the framework. For each tweet, we prompted ChatGPT-4o to classify the stance as pro-abortion, anti-abortion, or neutral, and to justify its classification based on the stance triangle using evidence from the language of the tweet.

In the final step of Phase 1, the emotion and stance codes assigned by the two human coders and ChatGPT were compiled into Excel spreadsheets. Each tweet was labeled for AFFECT (positive, negative, or neutral) and stance (pro, anti, or neutral). These categorical labels were then converted into numerical values (e.g., 1 = positive, 2 = negative, 3 = neutral for AFFECT; 1 = pro, 2 = anti, 3 = neutral for stance) to prepare the data for statistical analysis.

In Phase 2 of the study, we used a custom Python script to calculate inter-coder agreement between the human coders and ChatGPT. Cohen's Kappa (Cohen, 1960) was used to assess pairwise agreement (e.g., between ChatGPT and each human coder), as it corrects for chance agreement between two raters. Krippendorff's Alpha (Krippendorff, 2007) was used to handle multiple coders and varying levels of measurement and is therefore well suited for evaluating overall reliability across all coders. To explore patterns of agreement and disagreement, we also conducted chi-square tests to determine whether there were particular areas of significant differences in how human and AI coders analyzed AFFECT and stance.

In Phase 3, interpretations were made through consideration of the qualitative frameworks and coding practices, the inter-coder agreement metrics, and statistical comparisons to understand how consistently these were identified by human coders and ChatGPT.

Results

To address RQs 1 and 2, which examine the AFFECT (emotions: positive, negative, or neutral) expressed in U.S. abortion-related tweets from Boston and the stances (pro-abortion, anti-abortion, or neutral) taken in these tweets, we calculated the raw counts of emotion and stance as assessed by human coders and ChatGPT. For RQ3, we conducted reliability tests, and for RQ4, we performed chi-square tests on the emotion and stance-taking codes. The results for each analysis are presented below according to the corresponding research question.

RQ1: Emotions Expressed in Tweets

To address the first research question on the frequency of each emotion, we examined the number of positive, negative, and neutral codes assigned by each coder (Human 1, Human 2, and ChatGPT). Figure 13.5 presents the raw emotion counts from each coder's analysis.

Figure 13.5 shows that negative emotions were the most frequently observed, while positive and neutral emotions occurred at similar levels. It is worth noting that while the chart in Figure 13.5 shows similar counts of emotions, this does not mean that the same tweets were consistently classified with the same emotions by each coder. Even in cases where the same code was given by humans and ChatGPT, the reasoning for assigning the code may be different. To illustrate coder agreement and disagreement in emotion classification between human and AI coders, we provide two examples below. Example 1 in Figure 13.6 presents a tweet that all coders, human and AI, classified as neutral. Although they agreed on the emotion, their reasoning for this classification varied.

Figure 13.5

Frequency of Emotion Categories by Coder

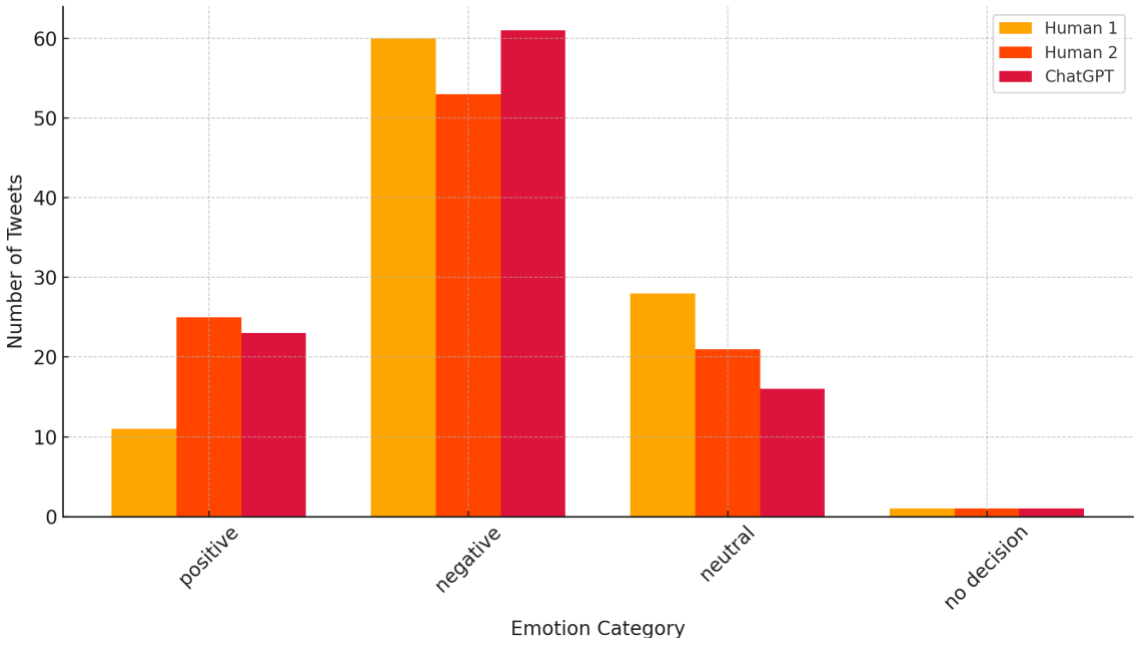
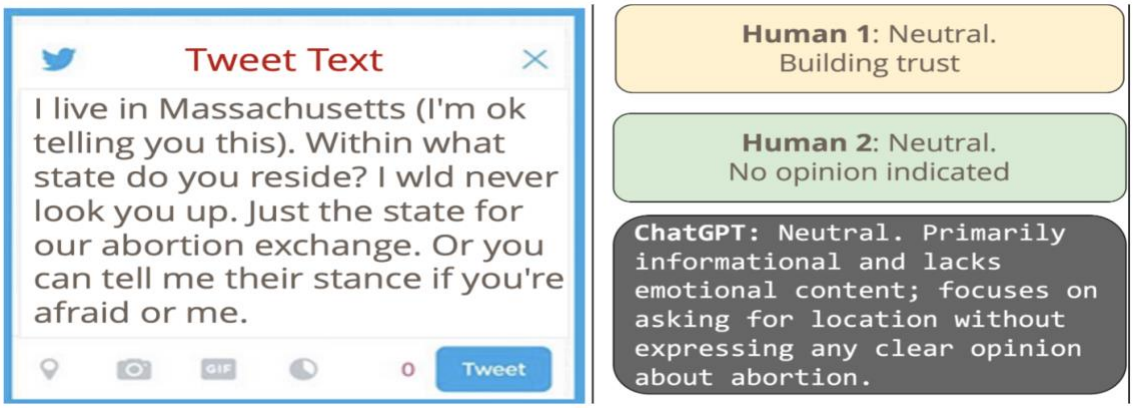


Figure 13.6

Example of Full Agreement on Emotion Classification across Coders



In contrast, Figure 13.7 shows a tweet where each coder identified a different emotion, illustrating how interpretations of the same content can vary significantly. This difference in coding highlights the challenges of achieving consistency in emotion analysis.

Figure 13.7

Example of Disagreement on Emotion Classification across Coders

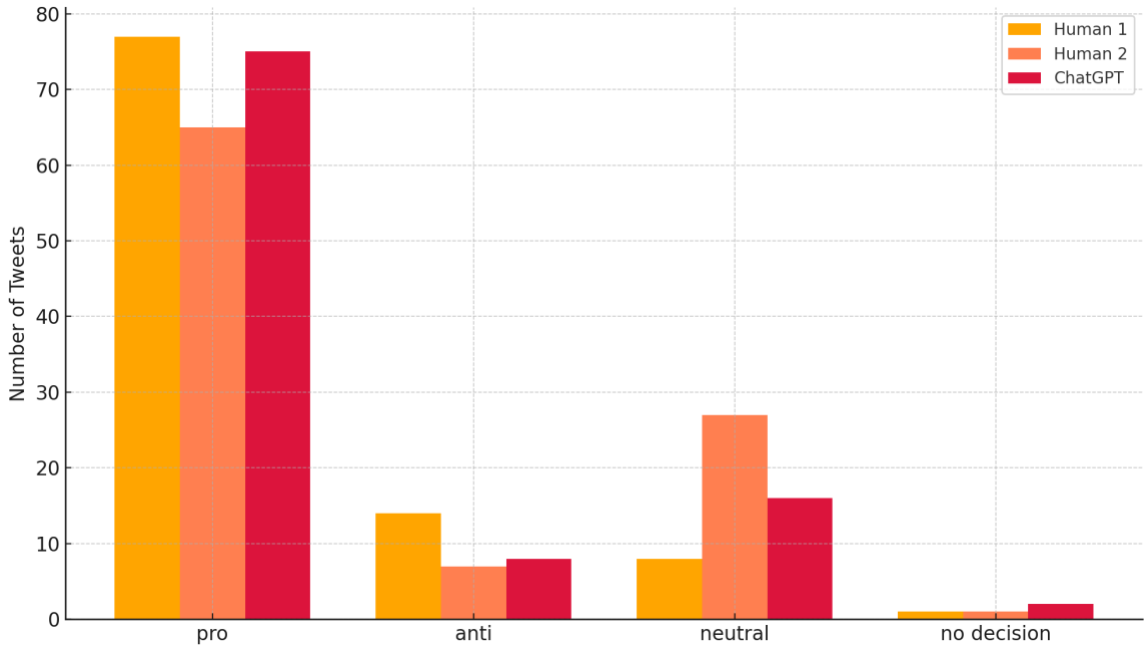


RQ2: Stance Expressed in Tweets

To address the second research question on the frequency of each stance, we counted the number of pro-abortion, anti-abortion, and neutral stance codes assigned by each coder (Human 1, Human 2, and ChatGPT). The chart in Figure 13.8 shows the raw stance counts from each coder's analysis and highlights the differences in stance analysis.

Figure 13.8

Frequency of Stance Categories by Coder

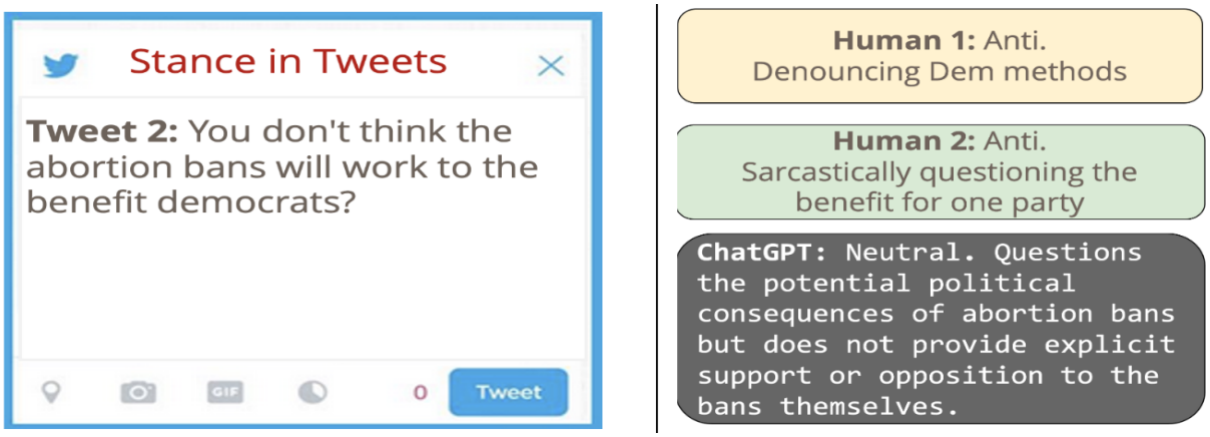


As can be seen from Figure 13.8, pro-abortion tweets are the most frequently observed, while anti-abortion and neutral tweets occur less frequently. Similar to emotion analysis, the chart in Figure 13.8 displays somewhat similar overall counts for each coder, but this does not imply that the same tweets were consistently classified with the same stance by each coder. These raw counts only reflect how many times each category was used by each coder; they do not indicate whether the same tweets were classified in the same way. To determine whether coders consistently assigned the same emotion and stance to the same tweets, we conducted chi-square tests to examine whether the distribution of emotion and stance labels differed significantly between coders.

The example we provide in Figure 13.9 is a tweet where both human coders agreed on the stance, but ChatGPT offered a different interpretation. This demonstrates how AI can sometimes miss the subtleties of stance-taking, even when human coders align in their analysis.

Figure 13.9

Example of Stance Disagreement between ChatGPT and Human Coders



RQ3: Agreement Across Coders

To address the third research question on coder agreement, we calculated inter-rater agreement between each pair of coders (Human 1, Human 2, and ChatGPT) for the emotion codes and for the stance codes using Cohen's Kappa. Overall reliability across coders was calculated using Krippendorff's Alpha.

Agreement in Emotion Analysis

The agreement in emotion analysis between each coder pair was found to be fair. As shown in Table 13.1, Kappa values for emotion analysis indicated fair agreement: 0.355 between Human 1 and Human 2, 0.283 between Human 1 and ChatGPT, and 0.322 between Human 2 and ChatGPT. With Kappa values between 0.21 and 0.40 generally considered fair (Landis & Koch, 1977), all coder pairs in this study demonstrated only fair agreement in identifying emotions.

Table 13.1

Results from Cohen's Kappa Analysis (N = 100 Tweets)

Pairs of Ratings	Cohen's kappa	SE	95% CI	
			Lower	Upper
Human 1 - Human 2	0.355	0.072	0.213	0.496
Human 1 - ChatGPT	0.283	0.079	0.128	0.439
Human 2 - ChatGPT	0.322	0.079	0.167	0.477
Average kappa	0.320			

Note. 100 subjects/items and 3 raters/measurements. Confidence intervals are asymptotic.

We also calculated Krippendorff's Alpha to assess overall agreement among the three raters (shown in Table 13.2). The Alpha value of 0.319 indicates fair agreement (Krippendorff, 2007), which means that the raters agreed more than by chance but with limited consistency. This result aligns with Cohen's Kappa findings, which also reflected fair agreement.

Table 13.2

Krippendorff's Alpha Results (N = 100 Tweets)

Method	Analysis	Krippendorff's alpha	SE	95% CI	
				Lower	Upper
Nominal	Emotion	0.319	0.054	0.202	0.414
	Stance	0.400	0.061	0.277	0.512

Note. 100 subjects/items and 3 raters/measurements.

Agreement in Stance Analysis

For stance-taking, Cohen's Kappa values in Table 13.3 indicate moderate agreement between Human 1 and Human 2 ($\kappa = 0.402$) and between Human 2 and ChatGPT ($\kappa = 0.468$), based on Landis and Koch's (1977) interpretive scale. The agreement between Human 1 and ChatGPT is fair ($\kappa = 0.335$), suggesting more limited consistency. According to these results, ChatGPT's stance coding aligned more closely with Human 2 than with Human 1. While kappa values in the moderate range suggest some degree of reliability, further refinement of prompts or coder guidelines may be needed before such AI-assisted annotation could be used in applied linguistics research contexts that demand higher levels of reliability.

Table 13.3

Cohen's Kappa Analysis for Stance (N=100 Tweets)

Pairs of Ratings	Cohen's kappa	SE	95% CI	
			Lower	Upper
Human 1 - Human 2	0.402	0.074	0.257	0.546
Human 1 - ChatGPT	0.335	0.092	0.155	0.515
Human 2 - ChatGPT	0.468	0.083	0.305	0.632
Average kappa	0.402			

Note. 100 subjects/items and 3 raters/measurements. Confidence intervals are asymptotic.

Table 13.2 shows the results of Krippendorff's Alpha calculated for the stance analysis, which yielded a value of 0.400, indicating moderate agreement (Krippendorff, 2007) among the three coders. While this suggests the raters agreed more than by chance, the agreement is not particularly strong and demonstrates both some consistency and notable differences in how stances were interpreted.

RQ4: Differences between Human Coders and ChatGPT

The inconsistency in the coding given by the two human coders across the complete sample posed a challenge for investigating the main purpose of the study, the discovery of ChatGPT's consistency with human coding in subjective aspects of language. The final research question, therefore, limited the sample of tweets to those receiving the same codes from the human coders. These tweets are considered clear cases, which would be expected to receive the same codes from ChatGPT. To address the final research question about whether there is a statistically significant difference between the human coders and ChatGPT in categorizing emotions and stances in the tweets, we conducted chi-square analyses.

Emotion Analysis between Human Coders and ChatGPT

To compare ChatGPT's emotion classifications with human coders, we conducted a Chi-square test using a subset of 62 tweets in which the human coders fully agreed on the emotion. Table 13.4 presents the contingency table of emotion classifications by ChatGPT versus human coders. Among the 62 tweets, 45 were labeled as negative by the human coders. ChatGPT only agreed with 35 of these, misclassifying the remaining 10 as either neutral (5) or positive (5) as shown in Table 13.4. This misalignment suggests that while ChatGPT was able to recognize a substantial portion of the negative tweets, it showed some inconsistency in differentiating between negative AFFECT and other categories. Agreement was stronger for positive tweets (6 out of 7 correctly identified) and mixed for neutral tweets (5 out of 10 matched).

Table 13.4

Chi-Square Contingency Table for Emotion Analysis

	ChatGPT			
Humans	Negative	Neutral	Positive	Total
Negative	35	5	5	45
Neutral	3	5	2	10
Positive	1	0	6	7
Total	39	10	13	62

Note. A Chi-square test revealed a statistically significant difference in classification patterns, $\chi^2(4, N = 62) = 30.96$, $p < .001$.

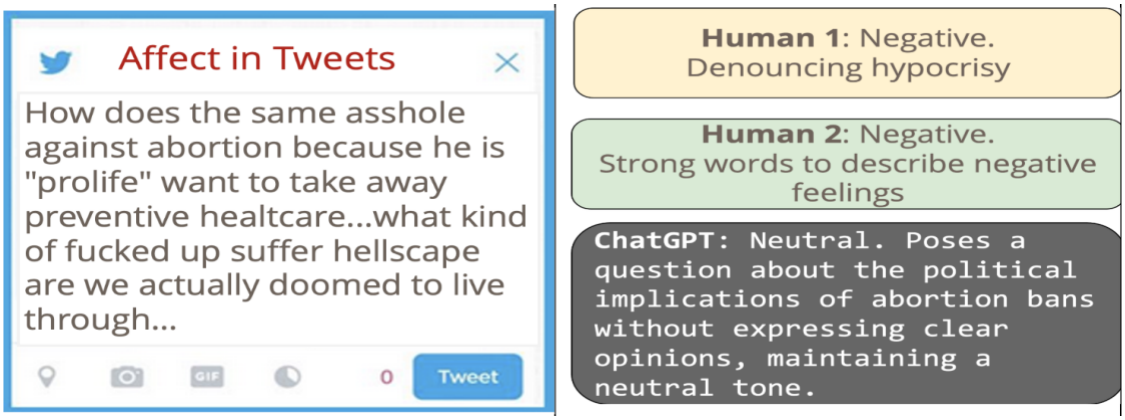
The Chi-square test yielded a statistically significant result $\chi^2(4, N = 62) = 30.963$, $p < .001$, which indicates that the differences in how emotion was classified by ChatGPT compared to human coders are unlikely to have occurred by chance. This suggests that the two sources did not categorize affective content similarly (Agresti, 2013). Thus, we conclude that ChatGPT's emotion classifications differ meaningfully and significantly from those of the human coders.

Below is an example tweet that both human coders categorized as negative due to its strong, vulgar language expressing clear dissatisfaction. However, ChatGPT classified the same tweet as neutral, overlooking the intensity of the language. This highlights a limitation in

ChatGPT's ability to fully capture the context and emotional weight of strongly worded expressions.

Figure 13.10

Example of Disagreement in Emotion Classification between Human Coders and ChatGPT



Stance Analysis between Human Coders and ChatGPT

To determine whether there were significant differences in how stance was classified by ChatGPT versus human coders, we conducted a Chi-square test using a subset of 70 tweets on which both human coders fully agreed. Table 13.5 presents the contingency table comparing ChatGPT's stance classifications to the consensus labels from human coders. Of the 70 tweets, 62 were labeled as pro-abortion by the human coders. ChatGPT correctly identified 58 of these, but misclassified the remaining 4 as either neutral (2) or anti-abortion (2). Among the four tweets labeled as anti-abortion by the humans, ChatGPT only identified 2 correctly and misclassified 1 as pro and 1 as neutral. For neutral tweets, ChatGPT matched only 2 of the 3, with the remaining one misclassified as pro-abortion. These results suggest that while ChatGPT was able to identify most pro-abortion tweets, its classifications for less common categories, such as neutral and anti-abortion stances, were less consistent with human judgments. This pattern may reflect challenges in recognizing stance cues that are more implicit or less frequently represented in the dataset.

Table 13.5

Chi-Square Contingency Tables for Stance Analysis

	ChatGPT				
Humans	Pro	Neutral	Anti	No Decision	Total
Pro	58	2	2	0	62
Neutral	1	2	0	0	3
Anti	1	1	2	0	4
No Decision	0	0	0	1	1
Total	60	5	4	1	70

Note. A Chi-square test revealed a statistically significant difference in classification patterns, $\chi^2(9, N = 70) = 105.68$, $p < .001$.

The Chi-square results confirm that the stance classifications produced by ChatGPT differ significantly from those of human coders. With a Chi-square value of 105.681 and a p-value below .001, these differences are unlikely to have occurred by chance. This indicates that ChatGPT's stance interpretations are not closely aligned with the human coders', particularly for categories with fewer instances, such as neutral and anti-abortion tweets.

The Chi-square results are illustrated by the following example, where a tweet classified as neutral by both human coders was identified as pro-abortion by ChatGPT, highlighting differences in stance interpretation. The human coders justified their categorization by noting that the tweet lacked language indicating pro or anti stance, while ChatGPT explained its pro-stance categorization as being based on "limiting the activities of pregnancy centers." This suggests that the human coders focused primarily on explicit language, whereas ChatGPT appeared to interpret the underlying implications of the text. This example highlights how ChatGPT's training data might lead it to engage in political interpretations.

Figure 13.11

Example of Disagreement in Stance Classification between Human Coders and ChatGPT



Discussion

The purpose of this study was to assess the use of a GenAI tool for discourse analysis of subjectivity in human language. Two aspects of subjectivity—emotions and stances—were coded by two human coders and by ChatGPT in U.S. abortion-related tweets from Boston, Massachusetts, posted between September 2022 and January 2023. Using Martin and White's (2005) Appraisal Framework, specifically the AFFECT subcategory, the coders attempted to identify positive, negative, and neutral emotions. Additionally, Du Bois' (2007) stance-taking framework was applied to classify tweets as pro-abortion, anti-abortion, or neutral. While ChatGPT demonstrated some agreement with human coders, particularly with Human 2 in stance

analysis, overall reliability and agreement were limited not only between ChatGPT and the human raters but also between the two human raters. This limited agreement highlights the inherent challenges in interpreting subjective content, particularly in emotionally charged and ideologically complex discourse.

Of particular interest were the subset of 62 and 70 tweets that received the same codes for emotion and stance, respectively, from both human raters. The Chi-square findings revealed significant differences in how ChatGPT and human coders classified emotions and stances, particularly in handling context-sensitive language. For example, ChatGPT often labeled emotionally charged or strongly worded tweets as neutral, even when human coders agreed they conveyed negative AFFECT. This pattern is evident in the emotion analysis contingency table (Table 13.4), where ChatGPT frequently misclassified tweets containing vulgar or intense language. These findings align with prior research (Dupré et al., 2020; Fatahi et al., 2024), which shows that human coders tend to outperform AI in interpreting spontaneous expressions, which can include strong language. However, it is also worth noting that interrater agreement between the human coders in our study was only fair to moderate, with Cohen's Kappa values of 0.355 for emotion and 0.402 for stance. Given these moderate levels of human agreement, ChatGPT's performance, while limited in some cases, did not fall substantially outside the range of variability observed between human coders. This suggests that although ChatGPT may struggle with context-dependent cues, its performance reflects challenges inherent in the task of classifying AFFECT and stance according to the frameworks used in this study.

Other studies have also reported similar outcomes regarding the challenges inherent in emotion classification. For example, Elyoseph et al. (2024) found that ChatGPT demonstrated greater emotional awareness than male participants in their study, while female participants slightly outperformed the model. Similarly, in our study, for emotion analysis, ChatGPT's classifications aligned more closely with the female coder ($\kappa = 0.322$) than with the male coder ($\kappa = 0.283$) and for stance analysis, agreement between ChatGPT and the female coder was higher ($\kappa = 0.468$) than with the male coder ($\kappa = 0.335$). These patterns suggest that ChatGPT's responses may be more in line with how some people interpret meaning based on implied or inferred content, rather than just the words that are explicitly stated. While ChatGPT may perform comparably to humans in some emotion recognition tasks, subtle emotional nuances continue to present challenges, as seen across both our study and prior research. One possible explanation for this difference lies in the methodological design: in our study, only two human coders were used, and their agreement was only fair. This suggests a need for more clearly defined coding criteria and improved rater training to enhance consistency in future studies. Establishing stronger guidelines and consensus-building practices would enhance the comparability and generalizability of findings in AI-human evaluation studies.

The challenges in accurately detecting emotions and subtle nuances, such as stance, in AI systems, as noted by Machová et al. (2024), were also evident in our study. ChatGPT's acknowledgment of its non-human nature when faced with complex content reflects an important self-awareness of its limitations. This aligns with our findings that current LLMs still struggle with detecting sarcasm and irony, which often rely heavily on context, tone, and cultural nuances. A final limitation we encountered was related to the Martin and White framework used for emotional analysis. While the framework is well-structured and offers clear categories for describing emotions, it doesn't come with a finalized or widely accepted list of words to guide

the analysis. Even Martin and White note that their 'maps of feeling' are more like hypotheses than fixed, evidence-based classifications. This lack of a concrete reference point makes it harder to apply the framework consistently, especially when comparing how humans and AI interpret emotional content. Similar challenges arose with the stance framework, which also lacked detailed operational guidance. In both cases, the frameworks offer valuable conceptual tools for analyzing subjectivity in discourse, but their practical implementation in this study was constrained by insufficient operationalization. For human coders, clearer annotation guidelines and consensus-building discussions may improve consistency. For AI systems, however, improvements may require task-specific fine-tuning or prompt engineering strategies aligned more closely with the theoretical underpinnings of the frameworks. These differences highlight the need to tailor operational approaches depending on whether the analysis is human- or AI-driven.

Conclusion

This study examined how ChatGPT compares to human coders in analyzing emotion and stance expressed in human language in politically sensitive tweets. Our findings highlight the complexity and subjectivity of such annotation, with interrater agreement among trained linguists being only fair. ChatGPT's classifications aligned more closely with one human coder than the other, suggesting that its output may reflect underlying biases from its training data, such as dominant discourse patterns and culturally specific interpretations. Results from the statistical analyses revealed significant differences between ChatGPT and human coders, particularly in detecting strong negative AFFECT and interpreting neutral stances. These findings highlight the need to test and adapt the linguistic frameworks when working with AI systems in qualitative analysis. In human-led discourse analysis, reliability is often strengthened through processes such as coder discussion and consolidation. To achieve similar reliability with AI, future research should explore methods for simulating or supporting such processes, such as adaptive prompting, iterative refinement, or structured human-AI interaction protocols.

References

Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons.

Ballesteros, J. A., Ramírez V, G. M., Moreira, F., Solano, A., & Pelaez, C. A. (2024). Facial emotion recognition through artificial intelligence. *Frontiers in Computer Science*, *6*, Article 1359471. https://doi.org/10.3389/fcomp.2024.1359471

Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., & Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, *95*, 96-108. https://doi.org/10.1016/j.techfore.2013.04.013

Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K. S., & Srinivasan, K. (2024). A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer Science Review*, *52*, Article 100632. https://doi.org/10.1016/j.cosrev.2024.100632

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage Publications.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), Article 100082. https://doi.org/10.1016/j.acorp.2023.100082

Du Bois, J. W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (pp. 139-182). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.07du

Dupré, D., Krumhuber, E. G., Küster, D., & McKeown, G. J. (2020). A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *Plos One*, *15*(4), Article e0231968. https://doi.org/10.1371/journal.pone.0231968

Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., & Hadar-Shoval, D. (2024). Capacity of generative AI to interpret human emotions from visual and textual data: Pilot evaluation study. *JMIR Mental Health*, *11*, Article e54369. https://doi.org/10.2196/54369

Fatahi, S., Vassileva, J., & Roy, C. K. (2024). Comparing emotions in ChatGPT answers and human answers to the coding questions on Stack Overflow. *Frontiers in Artificial Intelligence*, *7*, Article 1393903. https://doi.org/10.3389/frai.2024.1393903

Giaxoglou, K., & Spilioti, T. (2020). The shared story of #JeSuisAylan on Twitter: Story participation and stancetaking in visual small stories. *Pragmatics*, *30*(2), 277-302. https://doi.org/10.1075/prag.18057.gia

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), 1-3. https://doi.org/10.1073/pnas.2305016120

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating large language models in generating synthetic HCI research data: A case study. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* Article 433, 1-19. https://doi.org/10.1145/3544548.3580688

Hartson, H. R. (1998). Human–computer interaction: Interdisciplinary roots and trends. *Journal of Systems and Software*, *43*(2), 103-118. https://doi.org/10.1016/S0164-1212(98)10026-2

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81-88. https://doi.org/10.1080/02763869.2018.1404391

Huang, Z. Y., Chiang, C. C., Chen, J. H., Chen, Y. C., Chung, H. L., Cai, Y. P., & Hsu, H. C. (2023). A study on computer vision for facial emotion recognition. *Scientific Reports*, *13*(1), Article 8425. https://doi.org/10.1038/s41598-023-35446-4

Krippendorff, K. (2007). *Computing Krippendorff's alpha reliability*. ASC. https://perma.cc/YCU7-K3J8

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Li, M., & Conrad, F. (2024). *Advancing annotation of stance in social media posts: A comparative analysis of large language models and crowd sourcing*. ArXiv. https://doi.org/10.48550/arXiv.2406.07483

Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4-11. https://doi.org/10.1109/THFE2.1960.4503259

Machová, K., Szabóova, M., Paralič, J., & Mičko, J. (2023). Detection of emotion by text analysis using machine learning. *Frontiers in Psychology*, *14*, Article 1190326. https://doi.org/10.3389/fpsyg.2023.1190326

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). A dataset for detecting stance in tweets. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), *9*, 3945-3952. https://aclanthology.org/L16-1623/

Myers, B. A. (1998). A brief history of human-computer interaction technology. *Interactions*, *5*(2), 44-54. https://doi.org/10.1145/274430.274436

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., & Bosco, C. (2017). Hate speech annotation: Analysis of an Italian Twitter corpus. In R. Basili, M. Nissim, & G. Satta (Eds.), *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), 2006*, 1-6. https://hdl.handle.net/2318/1659197

Rahman, M. M., Terano, H. J., Rahman, M. N., Salamzadeh, A., & Rahaman, M. S. (2023). ChatGPT and academic research: A review and recommendations based on practical examples. *Journal of Education, Management and Development Studies*, *3*(1), 1-12. https://doi.org/10.52631/jemds.v3i1.175

Sharma, E., Saha, K., Ernala, S. K., Ghoshal, S., & De Choudhury, M. (2017). Analyzing ideological discourse on social media: A case study of the abortion debate. *Proceedings of the 2017 International Conference of the Computational Social Science Society of the Americas*, Article 3, 1-8. https://doi.org/10.1145/3145574.3145577

Taufek, T. E., Nor, N. F. M., Jaludin, A., Tiun, S., & Choy, L. K. (2021). Public perceptions on climate change: A sentiment analysis approach. *GEMA Online Journal of Language Studies*, *21*(4), 209-233. http://doi.org/10.17576/gema-2021-2104-11

Wikström, P. (2019). Acting out on Twitter: Affordances for animating reported speech in written computer-mediated communication. *Text & Talk*, *39*(1), 121-145. https://doi.org/10.1515/text-2018-2021

Zhang, B., Fu, X., Ding, D., Huang, H., Dai, G., Yin, N., Li, Y., & Jing, L. (2023). *Investigating chain-of-thought with ChatGPT for stance detection on social media*. ArXiv. https://doi.org/10.48550/arXiv.2304.03087

About the Authors

Sebnem Kurt (skurt@iastate.edu) holds a Ph.D. in the Applied Linguistics and Technology program at Iowa State University. She is a former Fulbright scholar and a TESOL certificate holder from Idaho State University. Her research interests include technology-assisted language learning, teaching and assessment, as well as online teacher training. Her most recent contributions to the Technology for Second Language Learning Conference (TSLL) can be found in Exploring Open AI in Language Teacher Professional Development: https://doi.org/10.31274/isudp.2024.154.12. Author ORCiD: https://orcid.org/0000-0003-4191-5277.

Mahdi Duris (mduris@iastate.edu) holds a Ph.D. in Applied Linguistics & Technology at Iowa State University, focusing on the use of technology in second language pronunciation, second language writing, and AI applications in education. His research interests center on pronunciation, phonetics, and phonology, especially regarding the automation of intelligibility ratings, as well as Automated Writing Evaluation (AWE) and the broader integration of AI in educational contexts. https://mahdiduris.com. Author ORCiD: https://orcid.org/0000-0002-1569-1655.

Chapter 14

ChatGPT, Second Language Pragmatics, and Critical AI Literacy

Robert Godwin-Jones
Virginia Commonwealth University

The ability of generative AI, such as ChatGPT, to produce language that closely resembles human-produced speech has led to claims that AI chatbots facilitate an authentic, interactional language learning environment. The assumption is that AI output is linguistically and culturally authentic enough that it could substitute in language learning settings for human interlocutors or for intercultural exchange. Such a view ignores how AI systems reproduce language and the limitations of that process for the linguistic features and the cultural content of the resulting output. Acknowledging those inadequacies is key to developing critical AI literacy, an essential component of becoming an informed AI user. The mathematical model of language in AI lacks the sociocultural grounding humans have through sensorimotor interactions and from simply living in the real world. Studies of AI's capabilities to engage in pragmatically effective language use have shown significant limitations. While AI systems can gain pragmalinguistic knowledge and learn appropriate formulaic sequences through the verbal exchanges in their training data (politeness conventions, for example), they have proven to be much less effective in sociopragmatic engagement, that is, in generating contextually acceptable speech reflecting an interlocutor's state of mind. AI systems will inevitably improve through larger datasets and integrated multimedia. However, those measures will not substitute for the human experience of negotiating common ground linguistically and culturally in social interactions, and therefore, the ability to deal with nuanced pragmatic scenarios. That fact has important repercussions for the use of AI chatbots in instructed second language acquisition.

Cite as: Godwin-Jones, R. (2025). ChatGPT, second language pragmatics, and critical AI literacy. In C. A. Chapelle, G. H. Beckett, & B. E. Gray (Eds.), *Researching generative AI in applied linguistics* (pp. 274-286). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2025.211.14
©2025 Godwin-Jones. Published under a CC BY license.

Introduction

Since its release in late 2022, the ability of ChatGPT to generate human-like language output has led to significant transformations in many areas of human activity. Along with other generative AI systems for text and image production, educational practices have been changed in myriad ways. For second language acquisition (SLA), AI offers the opportunity to engage in written and spoken exchanges, providing the means to practice the second language (L2) with an always available and patient partner. Moreover, the AI chatbot can be instructed to use the L2 at a specified level of proficiency, to take on a particular persona, as well as to provide corrective feedback. The ability of AI chatbots to engage in free-form conversation–on any topic, at considerable length, and in the context of previous exchanges in the chat– makes their performance radically different from previous rule-based chatbots that relied on scripts and keyword detection to generate responses.

The typically natural flow of conversation in dialogue with an AI system raises the important question addressed in this chapter, namely, whether the high-quality language generation from ChatGPT extends to a capacity to use language in socially and culturally appropriate ways. Studies examining AI's pragmatic abilities have shown serious deficiencies that derive in large part from the statistical (rather than social) process of how its language abilities are developed and from the culturally biased training data from which it generates output. While ChatGPT and other generative AI provide valuable opportunities for language learning, their use needs to be predicated on a clear understanding of how reliable they are likely to be in particular contexts of use. Their use calls for the development of critical AI literacy for both language learners and teachers (Darvin, 2025).

Unless indicated otherwise, the studies discussed in this chapter used ChatGPT 3.5 from OpenAI. "AI" as used here references text-based generative AI systems like ChatGPT, Gemini, or Claude. Pragmatic theories discussed here and applied to ChatGPT include the Grice Cooperative Principle (Grice, 1989), intercultural pragmatics (Kecskes, 2019; McConachy, 2019), and Wee's (2025) hearer-based pragmatics. It will be argued that while traditional pragmatic (Gricean) theories applied to ChatGPT demonstrate limited interactional skills, alternative approaches may be more appropriate when dealing with the novel situation of a machine approaching human communicative abilities. Of assistance in characterizing that new dynamic of machine-human interactions are theories associated with posthumanism, such as sociomaterialism (Guerrettaz et al., 2021). As part of the development of critical AI literacy, L2 learners will need to be aware of the limitations of AI chatbot use for the development of their pragmatic and interactional skills.

Linguistic and Cultural Authenticity

With the significant new machine-based ability to generate language that closely resembles human speech, it is important for L2 learners and teachers to know if interactions with AI can be reliably used in place of face-to-face interactions in the target language: "As humans engage with Artificial Intelligence (AI) more frequently, we need to know what exactly students are practicing in such technology-mediated tasks without 'real' human interlocutors in order to utilize the technology most effectively for L2 practice" (Dombi et al., 2024, p. 35). Some studies

of AI use for language learning have found that AI chatbots "facilitate an authentic, interactional language learning environment" (Chiu et al., 2023, p. 2) and that AI use is "essential for promoting cultural sensitivity, intercultural competency, and global awareness" (Anis, 2023, p. 64). The assumption is that AI output is linguistically and culturally authentic enough that it could substitute in language learning settings for human interlocutors. As AI becomes increasingly prevalent in educational, professional, and social settings, the issue of trustworthiness comes to the fore: To what extent can we view exchanges with AI chatbots as "authentic" despite the fact that those systems have developed their language abilities through statistical processes and without any world knowledge or interactions with humans?

The concept of authenticity in SLA is complex and has a long history of controversy over what qualifies as "authentic" and whether that actually matters for language learning (Gilmore, 2019). Certainly, no one would mistake rule-based chat systems for anything close to an "authentic" conversation partner. In contrast, AI chatbots are flexible and responsive; in some contexts (such as role play), they are able to be proactive in dialogues, initiating conversation topics. Yet that responsiveness is different in origin and nature from the dynamics of a human conversation. An AI response to a user utterance or to a prompt has a minimal social aspect (mostly limited to politeness forms). Rather than socially based, AI output is statistically generated, built on patterns in its target data and algorithms developed from the deep learning process in its layers of artificial neural networks.

The extensive data on which systems like ChatGPT are trained, including many samples of human speech in interactional situations, result in an ability to handle language used in conversational exchanges in conventional ways. That transactional language ability, however, does not necessarily translate into natural-sounding individual conversations, which involve much more than the use of stock phrases and ritualized language (Dombi et al., 2024). The mathematical model of language in AI lacks the sociocultural grounding humans have through sensorimotor interactions and from simply living in the real world. That fact necessarily colors the ability to use language in the way humans have been socialized into doing; thus, to characterize an AI conversation partner as linguistically "authentic" is questionable (Poole & Polio, 2023).

As might be expected, cultural authenticity and sensitivity on the part of AI are limited as well. ChatGPT's training data, based on a wide scraping of digital and online text, is culturally biased, as it contains predominantly Anglophone, Western-oriented sources, resulting in a bias in output towards a "WEIRD" perspective on the world (Western, Educated, Industrialized, Rich, Democratic; Atari et al., 2023), thus not reflecting the cultural beliefs, traditions, and practices of the majority of the world. Even when prompted to take a particular perspective aligned with an identity that is non-WEIRD, the output will be filtered through the system's inherent WEIRD biases and therefore will present a distorted and possibly prejudicial perspective (Stokel-Walker, 2024). That is particularly an issue for Indigenous cultures for which AI performance is typically either severely limited or non-existent (Godwin-Jones, 2025). The situation is not likely to improve as language models increase in size, as the data sources will not change significantly. The commercial entities owning the major AI system have little financial incentive to accommodate other languages and cultures. The cultural inauthenticity of ChatGPT and other systems necessitates an awareness that output is not neutral; like all technology tools, there are social and technical dimensions to its use (Darvin, 2023).

An additional concern culturally is the impression conveyed by the self-assured, confident tone of ChatGPT output: “The chatbot’s preferred use of directives as an engagement strategy further conveys an ethos of credibility and authority, encouraging users to readily accept its responses as trustworthy and uncritically follow its recommendations and suggestions” (Van Poucke, 2024, p. 15). The peremptory tone in AI output, along with perspectives shaped by dominant Anglocentric ideological positions, may impose a particular worldview on users, not conducive to creating an inclusive learning environment. The cultural biases inherent in AI, as well as the absence of social experience, limit their ability in many contexts (and in most languages other than English) to function reliably as full-fledged expert speakers and, therefore, as trustworthy chat partners.

AI’s Language Model and Pragmatics

Studies examining AI's abilities as a conversation partner have been approached from a variety of perspectives. One of those is from the viewpoint of the nature of human language, comparing theories of how language ability is developed. Generative linguistics posits an innate (and unique) human language ability. Wee (2025) points out that the Chomskyan postulation of a species-specific biological language capacity means that what is produced by AI “must be designated non-language, or such use is conceded to be language, which then begs the question of how such language might be possible when there presumably is no language faculty to license its manifestation” (p. 133). In contrast, usage-based language models posit that language is based on and shaped by actual language use in context (Ellis, 2019). Rather than seeing language as a set of rules and fixed syntactical structures, usage-based linguistics postulates that language consists of patterns that are used and reused, varying from single words to phrases and lexicogrammatical constructions. With repeated use, these structures become cognitively entrenched within a community of practice. From that perspective, language is fundamentally situated in recurring pragmatic patterns. Pragmatics, at the core of real-world language use, has no hard and fast rules but is learned through social and interactional experiences.

Clearly, ChatGPT does not have an innate language ability, nor does it have social experience to draw on. AI can be said to confirm a user-based approach to language in that its language abilities come from a mastery of patterns (Godwin-Jones, 2021). Through machine learning using a complex statistical and algorithmic process, speech is translated into mathematical symbols, and patterns in the data are uncovered. The resulting large language model (LLM) allows for the prediction of the next item in a string of text and extrapolating from there to sentences, paragraphs, and extended discourse. Through variable semantic weighting to words and phrases, ChatGPT has become much closer to sounding “natural” in language production than one might expect from a statistical process. That has been assisted through fine-tuning from reinforced learning with human feedback (RLHF).

While the use of repeated word constructions in human communication is based on frequency, the exact assemblage of elements is dependent on context, as many variables – social, cultural, situational–shape the dynamic nature of human communication. Meaning-making comes through verbal language but also by nonverbal means, gestures, tone of voice, body language, and paralanguage. The situated and contingent nature of language use highlights the importance of considering the social and cultural factors that go into the complex process of

interacting with another human. That process unfolds interactively through adjusting communicative processes to the dynamics of a conversation, feeling out common ground, and following socially determined interactional practices. The latter include conversation openings/closings, turns at talk, speech acts, topic management, repairs, adjacency pairs, and more (Dombi et al., 2024).

The statistical model of AI differs markedly from how humans acquire language, through a gradual process of socialization. Additionally, humans have recourse to nonverbal cues that AI lacks. Our multiple experiences from childhood on participating in verbal and nonverbal meaning-making add another dimension to our language competence, the ability to "read" our interlocutors, that is to be able to gauge the other's thoughts, intentions, and emotions, often referred to as "theory of mind" (also labeled "folk psychology" by philosophers of language; Frankish, 2024). That ability allows us to understand meaning when not literally encoded in language. ChatGPT's limited capacity in theory of mind (Brunet-Gouet et al., 2023) correlates with findings that explore its inability to infer meaning (Ruis et al., 2024). That is one of several areas of meaning-making in AI that have demonstrated limited pragmatic abilities, the capacity to use language appropriate to the social, cultural, and interactional contexts.

Studies in pragmatics often reference the Gricean Cooperative Principle (Grice, 1989). It postulates that in social contexts, humans strive to be cooperative in order to reach a mutually satisfactory outcome. To achieve that result, Gricean pragmatics asserts that certain maxims need to be followed. Speakers need to say what is needed for the context but nothing more (maxim of Quantity). Utterances need to be trustworthy (Quality) and pertinent to the conversation (Relation). Language needs to be unambiguous and clearly understandable (Manner). The Cooperative Principle and the accompanying maxims have been applied to ChatGPT and also to other AI systems. The consensus is that, judged from Gricean pragmatics, AI output does not conform. Quantity is particularly problematic, in that AI chatbots tend to be verbose and repetitive (Harnad, 2024). The RTHF process leads ChatGPT to always generate output, whether or not there is anything substantive to say; unlike in human conversation, silence is not an option (Lee, 2024). Its training instructs AI to be relentlessly helpful. Quality is clearly suspect because AI can "hallucinate", creating misinformation and outright fabrications (Giannakidou & Mari, 2024).

Following the maxims of Relation and Manner requires, on the one hand, the ability to interpret the conversational context and, on the other, to achieve some understanding of the partner's preferred conversation style (Bonard, 2024). Those are insights not within the mind-reading capabilities of AI (Butterfill, 2024). Put clearly by Tao et al. (2024), "ChatGPT does not observe the Gricean maxims of Quantity, Quality, Manner and Relevance" (p. 2). The authors of that study point out that ChatGPT is particularly deficient in grasping implied meanings: "Understanding intentions and deriving indirect meanings, such as conversational implicatures, are thus central to human communication, yet a difficult task for LLMs" (Tao et al., 2024, p. 2). Other studies have likewise pointed to those deficiencies (Kim et al., 2023; Qiu et al., 2023) as well as to other linguistic features that rely on non-literal formulations, such as irony, sarcasm, and humor.

Studies of AI performance in pragmatics typically use role play and discourse completion tasks to measure effectiveness, particularly for inferred meanings and non-literal language use.

Those are approaches frequently used in clinical settings with patients on the autistic spectrum or those with brain injuries. Researchers have found that they are often not able to use standard scenarios because ChatGPT's training data (through textbooks and research articles) contains the most frequently used examples, and therefore, the system knows the expected and appropriate responses (Lee, 2024). Using novel test examples, performance varied in studies, with results improving only if ChatGPT was given additional assistance in responding. In Chen et al. (2024), better performance was obtained by providing the AI systems with "attitudinal indicators" that were not included in the scenarios used by the human subjects. Those instructed the AI in what feelings humans would likely have in that context. After presenting a role play testing performance on apologies, the study provided this prompt to ChatGPT: "Write the conversation as if you feel apologetic and as if the classmate's friend feels annoyed" (p. 17). In a study using implied meanings in short dialogues, ChatGPT needed added hints to uncover the actual meaning (Brunet-Gouet et al., 2023).

Other studies administering benchmark tests of language ability to AI have found that a system would perform adequately on one set of data but then fail utterly when there were slight changes in the wording. Bonard (2024) attributes that to "shortcut learning", systems keeping rigidly to the parameters derived from training data rather than adapting output to the context. AI often performs, according to the author, "by detecting spuriously correlated keywords rather than with humanlike abilities" (Bonard, 2024, p. 22). Butterfill (2024) comments that AI systems may perform well on tests of their language and reasoning abilities from "sensitivity to distributional patterns" (p. 139), rather than anything resembling analytical thought or genuine insights.

Through its training data, ChatGPT has acquired those "distributional patterns," the language constructions to use in many pragmatic contexts, such as speech acts. In that way, AI systems are able to deal competently with pragmalinguistic forms, responding, for example, in appropriate terms for apologies or requests. They are less successful in sociopragmatic contexts, that is adjusting the language to the flow of a conversation in ways that are socially and culturally appropriate. In real-world contexts, conversants will have a range of responses that are not pre-determined but depend on speakers' interpretations of their partner's words but also of the meaning behind the words: "Casual conversation has a wide range of allowable and unpredictable contributions which are negotiated in the flow of communication" (Bardovi-Harlig, 2020, p. 55). Some of the likely variations are culturally determined, and others are highly personal. It is possible that a speaker may be aware of expected pragmatic norms but, for personal, interpersonal, or sociopolitical reasons, modify or outright reject expected pragmatic choices. Using AI for learning pragmatic norms might run counter to individual identity positioning, given that it will assume the mainstream values and behaviors it has learned from its training data.

AI chatbots use a one-size-fits-all approach when it comes to pragmatic language use, relying on their store of phrases, rather than adapting to the context and user. That limitation is important to realize in the implications for AI use as free conversation practice partners in L2 learning and teaching. Martinenghi et al. (2024) summarize the language abilities of AI systems in this way: "Large language models lack functional linguistic competence, that is the ability to rely on world knowledge to form models of the situation and the interlocutors and to engage in social pragmatic understanding of the communicative intentions" (p. 115). The "functional ability" referenced here includes interactional competence, engaging appropriately with the

interlocutor. AI functions adequately at transactional speech but often breaks down when it comes to relational discourse (Dombi et al., 2024), which is central to human communication.

Beyond Gricean Pragmatics

While traditional ways to measure pragmatic and interactional competence show substantial limitations in AI output, the idea has been advanced that the revolutionary nature of generative AI in reference to language generation may require new approaches (see multiple chapters in Strasser, 2024). Methods that seem well adapted to rule-based chatbots may not suffice for characterizing and understanding generative AI. Miehling et al. (2024) suggested adding two additional maxims to the Gricean list to account for the dynamics of AI, namely Benevolence (do no harm) and Transparency, the acknowledgement from AI of its limitations (for example, the ability to respond "I don't know"). Others argue that a Gricean approach is inadequate and new and different frameworks need to be applied (Bonard, 2024; Satyanarayan & Jones, 2024).

A possible direction to pursue in rethinking pragmatics to account for generative AI is to consider alternatives such as intercultural pragmatics (Kecskes, 2019; McConachy, 2019). Gricean pragmatics assumes that interactants are reliably trustworthy, but that may not necessarily be the case in human encounters; there may be deliberately misleading assertions or even nefarious intentions at play, as well as unintentional misrepresentations (Setlur & Tory, 2022). Assuming benevolence from an AI chatbot is currently problematic (Tao et al., 2024). Another assumption of the Grecian model is that those involved in conversation have *a priori* common ground based on shared cultural backgrounds and language practices. That is typically true for a conversation between speakers of the same first language (L1), but it cannot be assumed for conversants using an L2. Moreover, speakers of pluricentric languages (e.g., English, Spanish, Arabic) may differ markedly in respect to commonalities in language and culture. Speakers of language variants and dialects are likely to have varying experiences in finding common ground. Just because speakers may be using the same named language, one can not assume a single homogeneous speech community in terms of sociopragmatic behaviors (Bardovi-Harlig, 2020), a failing often seen in instructional pragmatics and in textbooks.

In contrast to encounters between speakers of the same L1, the conversations between speakers of different L1s, conversing in a common L2, may not begin on the same footing. Cultural values, beliefs, and behaviors may not mesh, as backgrounds and traditions will differ. Depending on proficiency levels, lingua franca interlocutors may not have recourse to the kinds of idiomatic, figurative language which characterizes the speech of L1 speakers. Studies have shown that in intercultural encounters in an L2, more literal and transparent language expressions are used (McConachy, 2019). As a consequence of lacking *a priori* common ground, lingua franca speakers will negotiate linguistic practices and cultural commonalities on the fly. Studies of English as a lingua franca have shown how that process frequently plays out, with speakers striving towards mutual understanding by engaging in cooperative behaviors (Baker & Sangiamchit, 2019).

Conversing with an AI chatbot can be seen as a kind of intercultural exchange in that there is no established common ground. Studies of interactions with bots have shown that humans tend to adjust their speech in ways that are similar to engaging in an exchange with a less-than-proficient speaker (Dombi et al., 2024). As in lingua franca exchanges, human speakers

tend to simplify their syntax, using shorter sentences as well as front-loading essential lexis (Dombi et al., 2024). Literal language is preferred. Repetitions are frequent. Conversational turns may unnaturally combine separate pragmatic functions into a single utterance, such as greeting, ritual “how are you,” and topic initiation (Dombi et al., 2024). Social niceties may not be observed (shortened openings and closing), but that may depend on whether and to what extent the synthetic partner is anthropomorphized (Wee, 2025). That relationship with AI likely depends on the extent of a user's experience with chatbots, as well as the user's purpose in engaging in dialogue. The stance of the human user will be determined by the flow of the conversation and the AI responsiveness, or lack thereof. A major difference from intercultural encounters is that the AI has been instructed to be responsive and cooperative, but that does not derive from an empathetic impulse or intent. AI responses to human users result from extracting and combining data from its training corpus, along with the reinforcement training it has received. The forms may be appropriate, but the intent and meaning behind them are empty.

Distributed Agency in Human-AI Interactions

The asymmetry of chatting with an AI differs from the co-constructed cooperative spirit that typifies human exchanges. AI has no life experience, mindreading capabilities, or visual cues to guide responses. In most contexts, the responsibility for guiding the conversation rests with the human user. In that sense, traditional pragmatics, which gives priority to speaker intent, does not align well with AI as a partner. Wee (2025) therefore advocates for a “hearer-based pragmatics” (p. 51) that better accommodates the different dynamics of machine-human interactions in the age of AI. The primacy of speaker intent in Gricean pragmatics, the author argues, is limiting, “privileging a culture-specific and anthropocentric view of individualized personhood, intention, and communication" (p. 4) that may not apply across cultures. The author demonstrates how the dynamic of speaker-hearer expectations often differs in non-Western (especially Indigenous) cultures. It certainly is true that the Gricean model has predominantly been applied to Anglophone contexts in monolingual settings (Kecskes, 2019) and in dyadic, face-to-face situations. Additionally, traditional pragmatics is logocentric, largely ignoring alternative meaning-making by way of visual cues, touch, or writing.

Rather than assigning a passive role to the hearer, Wee (2025) asserts, we should emphasize the hearer's essential role in attributing communicative intent to what is said, no matter what the speaker's intention may have been. The hearer draws on the theory of mind to determine if an utterance is meaningful and relevant. In the case of an exchange with an automated entity, the author argues, the hearer's (or reader's) interpretative process will determine to what extent that conversation should be seen as the equivalent of conversing with a fellow human, that is to what extent the machine is anthropomorphized. The hearer decides "what kinds of speech acts are being performed so as to generate the relevant implicature " (p. 16). In that scenario, “agency is distributed across assemblages of both humans and things" (p. 15). Although AI may not exhibit intentionality in the sense of conscious thought, that does not mean it lacks agency, that is, the ability to act and influence behaviors and events.

In fact, the exchange between AI and human user has been described as “synergistic collaboration” (Parker et al., 2024, p. 1) or “distributed agency” (Godwin-Jones, 2024, p. 5). As users try out different prompts to tailor output, a shared creative process emerges from the continuous give and take. That moves creativity (and authorship) away from being exclusively

human functions, a view that accords with posthumanist theories such as sociomaterialism (Guerrettaz et al., 2021). Sociomaterialism decenters humans as the sole sources of meaning-making, attributing that potential to objects and beings as well (Toohey & Smythe, 2022). Similar are Indigenous perspectives on the intertwined relationship between humans and their environment/tools, as familiar in the Global South (Canagarajah, 2021). In the context of AI use by humans, that involves granting greater recognition to the presence and participation of nonhuman interactants in communication, with varying degrees of autonomy and agency.

If we apply concepts of inclusiveness and circular causality to the relationship between AI and human users, it leads us to see a more nuanced and complex relationship. Language and agency cannot be considered as "abstract or essentialized phenomenon that exists independently of their material manifestation" (Wee, 2025, p. 139). The emphasis needs to be away from the primary positioning of the human speaker and towards relationships or *assemblages* without fixed boundary lines (Thorne et al., 2021). From that perspective, the interpretation of speech acts and implicature is a shared process between participants but also involves context widely conceived. Humans as hearers draw on the theory of mind to make meaning and seek relevance in what has been said, relying as well on the full range of symbiotic resources. AI, as we have seen, has a restricted set of receptors. Its world is one of vectors and algorithms, not of living organisms and ideas, and thus is in a sense parallel but separate from ours (Rust, 2024), yet the two intersect to create shared artifacts and experiences.

Conclusion

A further distancing between our human and AI worlds is the reality of the device and the interface we are using, which in and of themselves contribute to meaning-making (Kern, 2014). If we view conversing with AI as similar to intercultural encounters, there is, in a sense, double meditation by the use of the access device. These are factors to consider in analyzing the "naturalness" of the use of AI chatbots in SLA. To gain critical AI literacy, that understanding needs to be added to the awareness that the language knowledge of AI systems is based on statistical frequency, not human experience. As Jiang et al. (2024) comment, "critical AI literacy extends beyond mere technical proficiency with AI tools to encompass a deeper understanding of the ethical and socio-material implications of AI systems" (p. 924). The practical implications for use in language learning should be a critical awareness of the benefits and limitations of AI for different users and purposes. While AI chats can clearly be profitable for learners in practicing L2 forms, lexis, and connected discourse, the interactional and pragmatic shortcomings of AI systems call for caution in expectations of developing a naturalness of expression in the L2.

AI is in the process of rapid development so that enhancements such as multimodal integration, enhanced voice capabilities, computer vision, and additional human feedback will improve AI's pragmatic abilities. Those changes will likely lead AI to be more "social," which in turn will improve its usability for SLA, as well as for many other uses that AI agents will be put to in our everyday lives. The typically natural flow of conversation in dialogue with an AI chatbot raises the important question of whether that kind of exchange should be considered the equivalent of conversation with a fellow human. This is not just an issue for AI use in SLA, as AI bots and agents are quickly taking on social roles such as companionship, tutoring, and even therapy. We need to know if AI systems will be trustworthy, predictable, and benevolent. The

evolving nature of AI prevents a definitive response, but for use in SLA, ChatGPT and other AI systems should continue to be used, although with a clear understanding that they will only go so far in developing communication skills applicable to real-world encounters in social (non-transactional) contexts.

References

Anis, M. (2023). Leveraging artificial intelligence for inclusive English language teaching: Strategies and implications for learner diversity. *Journal of Multidisciplinary Educational Research*, *12*(6), 54–70. https://ijmer.s3.amazonaws.com/pdf/volume12/volume12-issue6(5)/9.pdf

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which humans?* PsyArXiv . https://doi.org/10.31234/osf.io/5b26t

Baker, W., & Sangiamchit, C. (2019). Transcultural communication: Language, communication and culture through English as a lingua franca in a social network community. *Language and Intercultural Communication, 19*(6), 471-487. https://doi.org/10.1080/14708477.2019.1606230

Bardovi-Harlig, K. (2020). Pedagogical linguistics: A view from L2 pragmatics. *Pedagogical Linguistics*, *1*(1), 44-65. https://doi.org/10.1075/pl.19013.bar

Bonard, C. (2024). Can AI and humans genuinely communicate? In A. Strasser (Ed.), *Anna's AI Anthology: How to live with smart machines?* (pp. 243-276). Xenomoi.

Brunet-Gouet, E., Vidal, N., & Roux, P. (2023). *Do conversational agents have a theory of mind? A single case study of ChatGPT with the hinting, false beliefs and false photographs, and strange stories paradigms*. Zenodo. https://doi.org/10.5281/zenodo.7637476

Butterfill, C. (2024). What mindreading reveals about the mental lives of machines. In A. Strasser (Ed.), *Anna's AI Anthology: How to live with smart machines?* (pp. 123-147). Xenemoi.

Canagarajah, S. (2021). Rethinking mobility and language: From the Global South. *The Modern Language Journal*, *105*(2), 570–582. https://doi.org/10.1111/modl.12726

Chen, X., Li, J., & Ye, Y. (2024). A feasibility study for the application of AI-generated conversations in pragmatic analysis. *Journal of Pragmatics, 223*, 14-30. https://doi.org/10.2139/ssrn.4545327

Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, Article 100118. https://doi.org/10.1016/j.caeai.2022.100118

Darvin, R. (2023). Sociotechnical structures, materialist semiotics, and online language learning. *Language Learning & Technology, 27*(2), 28–45. https://doi.org/10.64152/10125/73502

Darvin, R. (2025). The need for critical digital literacies in generative AI-mediated L2 writing. *Journal of Second Language Writing*, *67*, Article 101186. https://doi.org/10.1016/j.jslw.2025.101186

Dombi, J., Sydorenko, T., & Timpe-Laughlin, V. (2024). Openings and closings in human-human versus human-spoken dialogue system conversations. *Language Learning & Technology*, *28*(2), 32–61. https://doi.org/10.64152/10125/73571

Ellis, N. C. (2019). Essentials of a theory of language cognition. *The Modern Language Journal*, *103*, 39–60. https://doi.org/10.1111/modl.12532

Frankish, K. (2024). What are large language models doing? In A. Strasser (Ed.), *Anna's AI anthology: How to live with smart machines* (pp. 53–78). Xenomoi.

Giannakidou, A., & Mari, A. (2024). *The human and the mechanical: Logos, truthfulness, and ChatGPT*. ArXiv. https://doi.org/10.48550/arXiv.2402.01267

Gilmore, A. (2019). Materials and authenticity in language teaching. In S. Walsh & S. Mann (Eds.), *The Routledge handbook of English language teacher education* (pp. 299-318). Routledge. https://doi.org/10.4324/9781315659824-25

Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology, 25*(1), 4-19. https://doi.org/10.64152/10125/44747

Godwin-Jones, R. (2024). Distributed agency in language learning and teaching through generative AI. *Language Learning & Technology, 28*(2), 5-31. https://doi.org/10.64152/10125/73570

Godwin-Jones, R. (2025). Technology integration for less commonly taught languages: AI and pedagogical translanguaging. *Language Learning & Technology*, *29*(2), 11-34. https://doi.org/10.64152/10125/73609

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

Guerrettaz, A. M., Engman, M. M., & Matsumoto, Y. (2021). Empirically defining language learning and teaching materials in use through sociomaterial perspectives. *The Modern Language Journal*, *105*(S1), 3–20. https://doi.org/10.1111/modl.12691

Harnad, S. (2024). *Language writ large: LLMs, ChatGPT, grounding, meaning and understanding*. ArXiv. https://doi.org/10.48550/arXiv.2402.02243

Jiang, J., Vetter, M. A., & Lucia, B. (2024). Toward a 'more-than-digital' AI Literacy: Reimagining agency and authorship in the postdigital era with ChatGPT. *Postdigital Science and Education*, *6*, 922–939. https://doi.org/10.1007/s42438-024-00477-1

Kecskes, I. (2019). *English as a lingua franca: The pragmatic perspective*. Cambridge University Press. https://doi.org/10.1017/9781316217832

Kern, R. (2014). Technology as Pharmakon: The promise and perils of the Internet for foreign language education. *Modern Language Journal, 98*(1), 340–357. https://doi.org/10.1111/j.1540-4781.2014.12065.x

Kim, Z. M., Taylor, D. E., & Kang, D. (2023). *"Is the Pope Catholic?" Applying chain-of-thought reasoning to understanding conversational implicatures*. ArXiv. https://doi.org/10.48550/arXiv.2305.13826

Lee, B. J. (2024). Exploring the potential of AI for pragmatics instruction. *Technology in Language Teaching & Learning*, *6*(3), 1521-1521. https://doi.org/10.29140/tltl.v6n3.1521

Martinenghi, A., Donabauer, G., Amenta, S., Bursic, S., Giudici, M., Kruschwitz, U., Garzotto, F., & Ognibene, D. (2024). LLMs of Catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in board games' multi-party dialogues. *Proceedings of the 10th Workshop on Games and Natural Language Processing@ LREC-COLING 2024* (pp. 107-118). ACL. https://aclanthology.org/2024.games-1.12/

McConachy, T. (2019). L2 pragmatics as 'intercultural pragmatics': Probing sociopragmatic aspects of pragmatic awareness. *Journal of Pragmatics*, *151*, 167-176. https://doi.org/10.1016/j.pragma.2019.02.014

Miehling, E., Nagireddy, M., Sattigeri, P., Daly, E. M., Piorkowski, D., & Richards, J. T. (2024). *Language models in dialogue: Conversational maxims for human-AI interactions*. ArXiv. https://doi.org/10.48550/arXiv.2403.15115

Parker, J. L., Richard, V. M., Acabá, A., Escoffier, S., Flaherty, S., Jablonka, S., & Becker, K. P. (2024). Negotiating meaning with machines: AI's role in doctoral writing pedagogy. *International Journal of Artificial Intelligence in Education*, 1-21. https://doi.org/10.1007/s40593-024-00425-x

Poole, F. J., & Polio, C. (2023). From sci-fi to the classroom: Implications of AI in task-based writing. *TASK*, *3*(2), 243-272. https://doi.org/10.1075/task.00022.poo

Qiu, Z., Duan, X., & Cai, Z. G. (2023). *Pragmatic implicature processing in ChatGPT*. PsyArXiv. https://doi.org/10.31234/osf.io/qtbh9

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2024). *The Goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs*. ArXiv. https://doi.org/10.48550/arXiv.2210.14986

Rust, J. (2024). Minimal agency in living and artificial systems. In A. Strasser (Ed.), *Anna's AI Anthology: How to live with smart machines?* (pp. 93-122). Xenemoi.

Satyanarayan, A., & Jones, G. M. (2024). Intelligence as agency: Evaluating the capacity of generative ai to empower or constrain human action. In D. Huttenlocher & A. Ozdaglar (Eds.), *An MIT Exploration of Generative AI: From Novel Chemicals to Opera*. MIT Open Publishing. https://doi.org/10.21428/e4baedd9.2d7598a2

Setlur, V., & Tory, M. (2022). How do you converse with an analytical chatbot? Revisiting Gricean maxims for designing analytical conversational behavior. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-17). https://doi.org/10.1145/3491102.3501972

Stokel-Walker, C. (2024, March 8). AI chatbot models 'think' in English even when using other languages. *NewScientist*. https://www.newscientist.com/article/2420973-ai-chatbot-models-think-in-english-even-when-using-other-languages/

Strasser, A. (2024). *Anna's AI anthology: How to live with smart machines.* Xenemoi.

Tao, Y., Agrawal, A., Dombi, J., Sydorenko, T., & Lee, J. I. (2024). *ChatGPT role-play dataset: Analysis of user motives and model naturalness*. ArXiv. https://doi.org/10.48550/arXiv.2403.18121

Thorne, S. L., Hellermann, J., & Jakonen, T. (2021). Rewilding language education: Emergent assemblages and entangled actions. *The Modern Language Journal*, *105*(S1), 106–125. https://doi.org/10.1111/modl.12687

Toohey, K., & Smythe, S. (2022). A different difference in teacher education: Posthuman and decolonizing perspectives. *Language and Education*, *36*(2), 122–136. https://doi.org/10.1080/09500782.2021.1980002

Van Poucke, M. (2024). ChatGPT, the perfect virtual teaching assistant? Ideological bias in learner-chatbot interactions. *Computers and Composition*, *73*, Article 102871. https://doi.org/10.1016/j.compcom.2024.102871

Wee, L. (2025). *Automation in communication: The ideological implications of language machines*. Taylor & Francis. https://doi.org/10.4324/9781003467922

About the Author

Robert Godwin-Jones, Ph.D., is Professor in the School of World Studies at Virginia Commonwealth University. His research is principally in applied linguistics, in the areas of language learning and technology and intercultural communication. He writes a regular column for the journal *Language Learning & Technology* on emerging technologies.