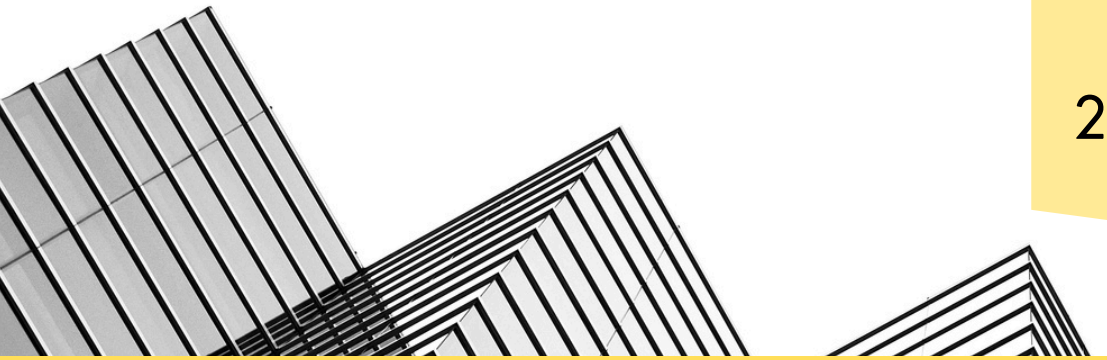
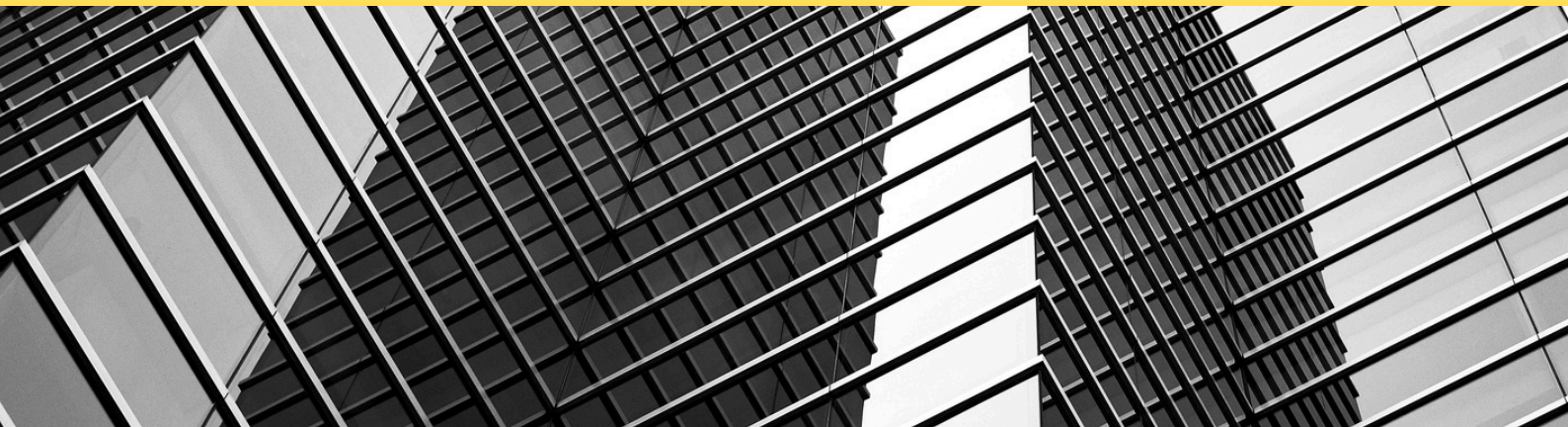
e-PROCEEDINGS

The 11th International Conference of the Asian Association for Language Assessment (AALA)

23-25 August 2025



Language Assessment in the Global South: Present Explorations & Future Directions



e-PROCEEDINGS

The 11th International Conference of the Asian Association for Language Assessment (AALA)

Language Assessment in the Global South: Present Explorations & Future Directions

23-25 August 2025

Published by:
Academy of Language Studies,
UiTM Cawangan Pulau Pinang

First Published 2025

e-ISBN 978-629-98082-2-0

Perpustakaan Negara Malaysia

Head Editor:
Dr. Norhaslinda Hassan

Editors:
Dr. Anwar Farhan Mohamad Marzaini, Muhammad Aiman Abdul Halim, Norhafizah Abdul Halil, Hanani Ahmad Zubir, Maizatul Akmal Mohd Mohzan, Muhammad Usamah Mohd Ridzuan

Published by:

Academy of Language Studies, UiTM Cawangan Pulau Pinang
04-382 3492
penang.uitm.edu.my
www.facebook.com/APBUiTMCPP

E-PROCEEDINGS THE 11TH INTERNATIONAL CONFERENCE OF THE ASIAN ASSOCIATION FOR LANGUAGE ASSESSMENT

e ISBN 978-629-98082-2-0

9 786299 808220

ACADEMY OF LANGUAGE STUDIES, UITM CAWANGAN PULAU PINANG
(online)

Preface by 11th International Conference of the AALA Local Chair

It is with great pleasure that I welcome you to the 11th International Annual Conference of the Asian Association for Language Assessment (AALA), hosted by Universiti Sains Malaysia (USM) in Penang. As the Local Chair, I am honoured to introduce this volume of conference proceedings, which reflects a wide range of scholarly and practical contributions under the theme "Language Assessment in the Global South." This year's theme invites critical reflection on the unique challenges, innovations, and perspectives emerging from regions often underrepresented in mainstream assessment discourse. It underscores the need to decolonise language assessment paradigms, foreground local knowledge, and address the socio-political realities that shape assessment practices in the Global South.

The papers included in these proceedings explore diverse issues—from equitable access to language testing and the use of local languages in assessment, to the integration of digital tools in resource-constrained environments. Collectively, they offer valuable insights into how educators, researchers, and policymakers across the Global South are reimagining assessment for greater fairness, relevance, and impact. The return of AALA to Malaysia is timely and meaningful, as it provides a platform to highlight regional voices and foster collaboration among language assessment communities in Asia, Africa, Latin America, and beyond. With over 300 participants and contributions from across the world, this conference serves as a vital space for dialogue, knowledge exchange, and critical inquiry.

I would like to express my sincere gratitude to the AALA Executive Committee for their trust in USM as host institution, and to our co-organizers from Universiti Utara Malaysia, Universiti Kebangsaan Malaysia, and Universiti Teknologi MARA for their unwavering support. Special thanks go to our conference sponsors, reviewers, keynote speakers, and the dedicated organizing committee whose hard work made this event and these proceedings possible. It is our hope that this collection of proceedings not only captures the scholarly spirit of AALA 2025 but also inspires continued work towards inclusive, context-sensitive language assessment practices that serve learners across the Global South and beyond.

Warm regards,

Dr. Alla Baksh bin Mohamed Ayub Khan
Local Chair
11th AALA International Conference
Universiti Sains Malaysia

AALA Conference 2025

Table of Contents

P004 Secondary Vocational Students' Perceptions of the English Classroom Assessment Environment and Assessment Task
He Yang[1], Yiran Miao[2], Sihui Yu[3] 1

P005 Evaluating the Role of Local Large Language Models in Open-Ended Writing Assessment: Innovations, Validity, and Ethical Considerations in a Canadian University Context
Johanathan Woodworth 9

P029 Exploring Rater Effects in Automated Assessment of EFL Learners' Paraphrasing Skills with NLP Metrics and Customized GPT
Minkyung Kim 19

P033 Language Assessment Knowledge of English Language Teachers in Assessing Four Language Skills
Nway Htway Khin 27

P038 Enhancing Assessment Literacy Among Educators: A Mixed-Methods Study on Policy Implementation and Professional Development Practices
Wu Xiaofan 34

P041 Gamification in Formative Assessment for Non-Majors: Engagement and Challenges
Thuy Duy T. Pham[1], Nha Phuong T. Nguyen[2] 43

P043 A Study on the Reliability of Scoring English Essays by Advanced English Learners Using ChatGPT
Jungyeon Koo 52

P047 A.I. or Human? A Study of AI-Powered Speech in Tertiary Level Listening Test for EFL Learners
*Khairi Fakhri Fazil[1], *Nur Ehsan Mohd Said[2], Pham Ngoc Bao Tram[3], Zeng Yijing[4]* 58

P049 Investigating Multiple-Choice Test Items Designed for Specific Reading Constructs: Insight From Item Writing
Pham Ngoc Bao Tram[1], Zeng Yijing[2], Khairi Fakhri Fazil[3], Nur Ehsan Mohd Said[4] 77

P053 TOEFL Multi-Faceted Validity in an Indonesian Context: A Systematic Review
Dian Purwitasari 88

P061 Investigating the Washback Effect of Introductory Arabic Language Writing Assessment on Learning
Ainul Rasyiqah Sazali 99

P080 Can AI Replace Human Raters? A Multi-Dimensional Analysis of AES Reliability Using GPT-4 and Beyond
Dai yi[1], Du Meirong[1], Wang Fan[2], Lu Min[3], Zhang Yu[4] 108

P087 Assessment Tasks and Autonomous Learning Motivation in Secondary Vocational English Classrooms: An Analysis Based on Structural Equation Modeling
He Yang[1], Sihui Yu[2], Yiran Miao[3] 115

P092 Psychological Factors Influencing College English Learning: An Investigation of Learning Interest, Learning Motivation, and Interaction Anxiety among Chinese University Students
Miao Yiran[1], Cheng Liying[2] And Yang He[3] 127

P099 CIPP Model-Based Evaluation and Optimization of Sino-Foreign Cooperative English Curriculum Systems
Ning Song[1], Jingyi Hu[2] 137

P106 Navigating Topic Familiarity: Malaysian MUET Candidates' Challenges and Strategies in a High-Stakes Speaking Assessment
**Nurul Iman Ahmad Bukhari[1], Tengku Mohd Farid Tengku Abdul Aziz[2], Atirah Izzah Che Abas[3], Arifuddin Abdullah[4], Alla Baksh Mohamed Ayub Khan[5]* 147

R002 The Application of Generative AI in Formative Assessment for English Writing Skills in K-12 Education: Challenges and Opportunities
Zihan Sun 154

R006 The Influence of AI-driven Writing Assistants on Students' Attitudes towards Writing Skills and Academic Honesty
Nur Ain Amani Mohd Azmi 157

P105 From Grammar Checks to Idea Support: Student Insights and Patterns of AI Use in English Assessment at a Malaysian Polytechnic.
Noor Darliza Binti Mohamad Zamri 167

P004

Secondary Vocational Students' Perceptions of the English Classroom Assessment Environment and Assessment Task

***He Yang[1], Yiran Miao[2], Sihui Yu[3]**

[1]School of Education, City University of Macau, Macau.[2]College of Foreign Languages, Chengdu University of Information Technology, China. [3]Mental Health Education and Counseling Center, Shenzhen Polytechnic University, China.
(E-mail: [1]M23092100755@cityu.edu.mo, [2]myr@cuit.edu.cn, [3]yusihui693380@126.com)
**Corresponding author: [1]M23092100755@cityu.edu.mo*

Abstract

This study investigates the relationship between assessment tasks (CATs) and the classroom assessment environment (CAE) as perceived by secondary vocational school students (SVSSs) in English classrooms. Previous research has highlighted the importance of integrating self-regulated learning (SRL) with classroom assessment, but there is a lack of studies focusing on vocational English classrooms. The aim of this study is to establish and test an analytical model that explains the relationship between SVSSs' perceptions of CATs and the CAE. A questionnaire adapted from the PATI and PCAES was administered to 281 SVSSs from a nationally key vocational school in Shanghai. Data analysis included validity and reliability tests, descriptive statistics, and stepwise multiple regression analysis. The results show that congruence with planned learning and authenticity have a significant positive predictive effect on a learning-oriented CAE, while transparency and student consultation positively predict a performance-oriented CAE. Authenticity, however, negatively predicts a performance-oriented CAE. This study enhances the applicability of CAE theory to vocational education contexts and provides evidence-based support for designing assessments that foster a learning-supportive environment in vocational English classrooms.

Keywords: English classroom; secondary vocational students; assessment tasks; assessment environment

1. Introduction

Current research emphasizes integrating self-regulated learning (SRL) with classroom assessment (CA) (Brandmo et al., 2020). The classroom environment critically shapes SRL development (Cleary et al., 2004; Cai et al., 2022), with the classroom assessment environment (CAE) serving as a tool to foster self-regulated learners (Brookhart, 1997; Brookhart & DeVoge, 1999). Within this framework, two key concepts merit attention: classroom assessment tasks

e-ISBN: 978-629-98082-2-0

(CATs) implicitly communicate educational priorities through evaluation, while CAE—formed by teacher practices and student interpretations—determines whether learners pursue mastery or performance goals (Stiggins et al., 2005; Alkharusi, 2011). Students' task perceptions (e.g., orientation and value) and collaborative learning environment perceptions significantly influence SRL (Dent et al., 2015; Cai et al., 2022), with teacher practices playing a pivotal role (Velayutham et al., 2013).

The classroom CAE theory (Brookhart, 1997) suggests CAE and CATs enhance learning motivation and achievement, with self-efficacy as a pivotal mediator: (1) CAE traits influence self-efficacy (Alkharusi, 2007,2009); (2) Self-efficacy reciprocally shapes CAE perceptions (e.g., higher efficacy links to greater motivation, Alkharusi, 2010); (3)Students' CATs perceptions (e.g., difficulty/scoring clarity) affect efficacy development (McMillan & Workman, 1998); (4) CATs authenticity is a key predictor (Van Dinther et al., 2014)—when perceived as authentic/transparent, students develop stronger subject-specific efficacy (Alkharusi, 2013).

Although CATs and CAEs are interrelated (Brookhart, 1997), research on their relationship remains insufficient (Alkharusi et al., 2014; Cheng et al., 2015). Recent studies based on self-regulated learning (SRL) demonstrate a bidirectional influence between them (Larenas et al., 2021; Zhang & Li, 2023), with variations observed across different student populations, necessitating broader sampling for further investigation. Research in the Middle East (Alkharusi et al., 2014) found that among male students, a learning-oriented CAE was associated with high levels of alignment with learning objectives, authenticity, student engagement, and diversity, whereas a performance-oriented environment correlated with high authenticity, engagement, and diversity but low alignment and transparency. Among female students, the learning-oriented pattern was similar, but the performance-oriented environment was linked only to high engagement, diversity, and low alignment. In China, research (Cheng et al., 2015) indicated that a learning-oriented CAE could be predicted by alignment with learning objectives, authenticity, student engagement, and transparency, while a performance-oriented environment was predicted by alignment, diversity, and engagement.

Compared with international studies, Chinese high-quality researches on English classroom assessment have rarely addressed the primary and secondary school levels (Jin & Sun, 2020). Among the limited number of studies, even fewer have focused on vocational English classrooms, and none have explored the relationship between CATs and the CAE in this context. The purpose of this study is to develop an analytical model that describes the multivariate relationship between vocational English students' perceptions of CATs and the CAE. This model can identify which perceptions of CATs predict specific CAEs. To achieve this objective, the study is guided by the following research questions. How do vocational English students perceive classroom CATs and the CAE? What is the relationship between CATs and the CAE?

2. Methods

2.1 Participants

Based on convenience sampling, 313 vocational students from a nationally key vocational school in Shanghai were selected—all of whom had attended open English classes (at or above the school level). The rationale for selecting students who had participated in open English

e-ISBN: 978-629-98082-2-0

classes was that such lessons typically involve higher-quality instructional preparation, ensuring the intentional design of CATs, which facilitates the collection of data on students' perceptions of CAT characteristics. A total of 281 valid questionnaires were collected, yielding an effective response rate of 89.78%. Among these 281 students, approximately two-thirds were male, 37.4% were from science and engineering disciplines.

2.2 Procedures

Data collection was conducted during a class meeting immediately after the open English lesson. Each participating student received a self-report questionnaire consisting of three sections. Students were informed that they were participating in a study investigating their perceptions of classroom CATs. Participation in the questionnaire was voluntary. Their responses would remain confidential. The three sections of the questionnaire were as follows. Demographic Information: including gender, students' major, class type and self-reported academic ranking within their class. Perceptions of CATs and the CAE: students were asked to evaluate these perceptions based on the open English lesson they had just attended. Under the guidance of their English teacher or homeroom teacher, students completed the electronic questionnaire via mobile devices in approximately 20 minutes.

2.3 Instruments

Two primary scales were employed to examine students' perceptions of English classroom CAT characteristics and the CAE. A 7-point Likert scale was utilized, with response options ranging from 1 (strongly disagree) to 7 (strongly agree). The specific measurement tools for each variable are described below.

CATs. This variable was measured using items adapted from the five dimensions (35 items total) of the English version of the Perceptions of CATs Inventory (PATI) developed by Dorman and Knightley (2006). The items assessed alignment with planned learning (e.g., I was assessed based on what the teacher taught me.), authenticity (e.g., The CATs in my English class were meaningful.), student consultation (e.g., I was asked what types of assessments I preferred in this course.), transparency (e.g., I was informed in advance about upcoming assessments.) and diversity (e.g., I had choices in CATs.). Scores for each of the five subscales were derived by calculating the mean of participants' responses, with higher average scores indicating more positive student perceptions of CATs.

CAE. This variable was measured using items adapted from the English version of the Perceived CAE Scale (PCAES) (Alkharusi, 2011), consisting of 16 items across two subscales. Learning-oriented CAE subscale (9 items): focused on whether classroom assessments facilitated student learning and mastery of course content (e.g., Assignments and tests encouraged thinking.). Performance-oriented CAE subscale (7 items), focused on assessment rigor, grading practices, public evaluation, and recognition (e.g., Exams and assignments were challenging for us.).

2.4 Data Analysis

Firstly, missing value check. Prior to conducting formal data analysis, the questionnaire data for missing values should be examined. Any questionnaire with incomplete responses or invalid answers (e.g., selecting multiple options for a single item) was excluded from the dataset. This process resulted in the removal of 32 invalid questionnaires, retaining 281 valid responses for

e-ISBN: 978-629-98082-2-0

subsequent analysis. Secondly, examination of validity and reliability. I conducted exploratory factor analysis (EFA) on both subscales to verify their underlying factor structures. Additionally, I assessed the internal consistency reliability of these scales. Thirdly, descriptive statistics. Descriptive statistical methods including means, standard deviations, and Pearson correlation analysis were used. Finally, stepwise multiple regression analysis to investigate the relationships between CATs and CAE. All statistical analyses were conducted using SPSS 26.

3. Results and Discussion

3.1 Validity and reliability

CATs Scale. Principal axis factoring (PAF) was employed for factor extraction, followed by direct oblimin rotation (an oblique rotation method). Items with factor loadings below 0.4 or cross-loading on two or more dimensions (with loadings ≥0.4) were removed. Specifically, the following items were deleted: items 1, 2, and 7 of authenticity; items 1, 2, and 7 of transparency; items 4, 6, and 7 of diversity.After EFA, the final structure consisted of five factors with 26 items, each factor containing at least four items. The factor loadings ranged from 0.53 to 0.97, and the cumulative explained variance reached 82.88%.The scale demonstrated excellent internal consistency: overall Cronbach's α is 0.94 and subscale Cronbach's α coefficients ranged from 0.86 to 0.97.

CAE Scale. Principal component analysis (PCA) was applied, followed by varimax rotation (an orthogonal rotation method). The EFA results yielded a two-factor structure with 16 items, each factor containing at least four items. The factor loadings ranged from 0.50 to 0.95, with a cumulative explained variance of 72.69%.The scale exhibited strong reliability: overall Cronbach's α is 0.86 and subscale Cronbach's α coefficients ranged from 0.92 to 0.95. Both scales demonstrated good internal consistency, supporting their validity for further statistical analysis.

3.2 Descriptive

The descriptive statistics are presented in Table 1. About CATs, in terms of mean scores, participants held the most positive views on the authenticity of CATs, followed by their alignment with planned learning. Overall, participants perceived all five CAT characteristics positively in English open classes. About CAE, participants reported positive perceptions of the learning-oriented CAE but less favorable views of the performance-oriented CAE. This suggests that students generally perceived English open classes as focused on learning enhancement rather than mere grade achievement.

The correlation analysis revealed the following key findings. Firstly, among CATs, all perceived dimensions showed statistically significant positive correlations.Strong correlations were observed between alignment with planned learning and authenticity. Weak correlations were found between alignment with planned learning and transparency; authenticity and transparency. Secondly, between CATs and learning-oriented environment, alignment with planned learning, authenticity, student consultation, and diversity showed significant positive correlations with the learning-oriented CAE. Among these, strong correlations were observed for alignment with planned learning and authenticity. No significant correlation was found with transparency. Thirdly, between CATs and performance-oriented environment, student consultation, transparency, and diversity exhibited significant positive correlations with the

e-ISBN: 978-629-98082-2-0

performance-oriented CAE. Among these, a moderate correlation was observed for transparency. No significant correlations were found for alignment with planned learning and authenticity.

Finally, between the two CAEs, no statistically significant correlation was found between the them.

Table 1 Descriptive

	cong	auth	cons	tran	dive	ATLE	MEAN	SD.
cong	--						6.35	0.98
auth	.80**	--					6.45	0.92
cons	.49**	.48**	--				5.57	1.63
tran	.18**	.14*	.51**	--			4.67	2.06
dive	.53**	.55**	.60**	.41**	--		5.98	1.25
ATLE	.67**	.74**	.38**	.08	.45**	--	6.53	0.77
ATPE	.00	-.07	.28**	.48**	.17**	-.04	3.27	2.10

Note. *p<0.05, **p<0.01. cong= alignment with planned learning, auth= authenticity；cons= student consultation, tran= transparency, dive= diversity, ATLE=learning-oriented CAE, ATPE= performance-oriented CAE.

3.3 Regression Analysis Results

As presented in Table 2, with the learning-oriented CAE as the dependent variable, two predictors—authenticity and alignment with planned learning—were retained in the model, while the other three independent variables were removed. These retained predictors demonstrated significant positive effects, collectively accounting for 55.6% of the variance (R^2 = .556). The β weights and significance levels were: authenticity (β = .55, p < .001) and alignment with planned learning (β = .23, p < .001). This indicates that both authenticity and alignment positively contribute to a learning-oriented CAE.

Table 3 shows results for the performance-oriented CAE as the dependent variable. Three predictors—transparency, authenticity, and student consultation—remained in the model, explaining 24.9% of the variance (R^2 = .249). Statistically significant effects emerged: transparency (β = .43, p < .001) and student consultation (β = .14, p < .05) exhibited positive predictive effects. Authenticity (β = -.19, p < .05) showed a negative predictive effect. These results suggest that transparency and student consultation facilitate a performance-oriented environment, whereas task authenticity inhibits its development.

Table 2 Regression analysis for the learning-oriented CAE

Variable	β	t	R^2	Adj.R^2	ΔR^2	F
Constant		10.908**				
auth	.55	8.323**				
cong	.23	3.415**				
			.559	.556	.019	176.045**

Note. **p<0.01. auth= authenticity, cong= alignment with planned learning.

Table 3 Regression analysis for the performance -oriented CAE

e-ISBN: 978-629-98082-2-0

Variable	β	t	R^2	Adj.R^2	ΔR^2	F
Constant		3.868**				
trans	.43	7.148**				
auth	-.19	-3.240*				
cons	.14	2.053*				
			.257	.249	.011	31.864**

Note. *p<0.05, **p<0.01. trans= transparency, auth= authenticity, cons= student consultation.

3.4 Discussion

Firstly, about the dual impact of authenticity, the findings of this study reveal that in secondary vocational English classrooms, students' perceived authenticity of CATs serves as a positive predictor of a learning-oriented CAE and a negative predictor of a performance-oriented CAE. This result partially aligns with Cheng et al. (2015). Authentic CATs enhance student motivation (Colthorpe et al., 2021). When students perceive tasks as highly authentic, they develop stronger self-efficacy beliefs in their learning capabilities (Alkharusi, 2013). Form more positive attitudes toward the interest, usefulness, and importance of learning materials. About contradictory finding on performance-oriented CAE, Alkharusi et al. (2014) found that performance-oriented environments may correlate positively with high authenticity. In achievement-driven settings, educational goals are complex, and authentic assessments may stimulate students' desire to demonstrate competence (both individually and normatively).

Secondly, alignment with planned learning enhances learning-oriented CAE. This study confirms prior research (Alkharusi et al., 2014; Cheng et al., 2015) showing that when students perceive assessments as aligned with learning objectives and personally meaningful, they develop stronger intrinsic motivation (I want to learn well.) and adopt more active learning strategies (I can learn well.) (Lizzio & Wilson, 2013). This alignment indirectly improves academic performance by encouraging deeper learning strategies and strengthening the link between assessment and long-term knowledge retention.

Third, the positive correlation between student consultation and a performance-oriented CAE aligns with prior findings (Alkharusi et al., 2014; Cheng et al., 2015). This phenomenon can be interpreted through instructional practices in secondary vocational English contexts. During demonstration lessons (including preparatory phases), teachers often reinforce their dominant role by treating graded CATs (e.g., exercises, tests) as tools to direct student learning. Students are primarily informed about test content and formats and examination question types. Given that secondary vocational students prioritize exam outcomes—viewing passing scores as sufficient learning achievement—heightened student consultation may inadvertently reinforce test-focused mindsets, thereby fostering a performance-oriented CAE.

Fourthly, Contrary to Alkharusi et al.'s (2014) negative/null correlations, I found a positive CAT transparency-performance orientation link in vocational EFL classes. This paradox likely stems from demonstration lessons' performative nature: teacher-centric assessments (e.g., score-focused criteria) make students mechanically rehearse CATs, reinforcing correct-answers-as-goal mindsets that amplify performance-oriented environments.

e-ISBN: 978-629-98082-2-0

4.Conclusion

This study focuses on English classes in secondary vocational schools, building upon CAE theory to develop and test an analytical model that explains how students' perceptions of CAT characteristics influence their perceptions of the CAE. Unlike prior research conducted in Western sociocultural and educational settings, this study establishes and empirically validates a theoretical model to examine whether perceived CAT characteristics can predict the CAE in Chinese vocational English classrooms. By doing so, it enhances the applicability of CAE theory and expands its scope to vocational education contexts. The study provides evidence-based support for the proposed model, reinforcing the theory's relevance in diverse educational systems. The findings highlight key relationships between daily classroom assessment activities, offering guidance for vocational English teachers on how to design assessments that foster a learning-supportive environment. Educators can use these insights to optimize CATs to enhance student engagement and create CAEs that prioritize learning over grades.

Limitations and Future Research Directions. On the one hand, as a descriptive, cross-sectional study, this research cannot establish causality between CATs and the CAE. Future studies should employ experimental or longitudinal methods to verify causal relationships among variables. On the other hand, the study used convenience sampling, recruiting participants from only one vocational school in Shanghai who had taken English open classes. Thus, the findings may not generalize to other regions of China. Future research should include a more representative sample from different provinces and educational backgrounds to strengthen external validity.

References

Alkharusi, H. A. (2007). *Effects of teachers' assessment practices on ninth grade students' perceptions of classroom assessment environment and achievement goal orientations in muscat science classrooms in the sultanate of Oman*. Kent:Kent State University.

Alkharusi, H. (2009). Classroom assessment environment, self-efficacy, and mastery goal orientation: A causal model. *INTI Journal: Special Issue on Teaching and Learning*, 104-116.

Alkharusi, H. (2010). A multilevel linear model of teachers' assessment practices and students' perceptions of the classroom assessment environment. *Procedia-Social and Behavioral Sciences*, 5, 5-11.

Alkharusi, H. (2011). Development and datametric properties of a scale measuring students' perceptions of the classroom assessment environment. *International Journal of Instruction*, 4(1), 105-120.

Alkharusi, H. (2013). Canonical correlational models of students' perceptions of assessment tasks,2015 motivational orientations, and learning strategies. *International Journal of Instruction*, 6(1),21-38.

Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2014). Modeling the relationship between perceptions of assessment tasks and classroom assessment environment as a function of gender. *The Asia-Pacific Education Researcher*, 23, 93-104.

Brandmo, C., Panadero, E., & Hopfenbeck, T. N. (2020). Bridging classroom assessment and self-regulated learning. *Assessment in Education: Principles, Policy & Practice*, 27(4), 319-331.

e-ISBN: 978-629-98082-2-0

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied measurement in education*, 10(2), 161-180.

Brookhart, S. M., & DeVoge, J. G. (1999). Testing a theory about the role of classroom assessment in student motivation and achievement. *Applied measurement in education*, 12(4), 409-425.

Cai, J., Wen, Q., Jaime, I., Cai, L., & Lombaerts, K. (2022). From classroom learning environments to self-regulation: The mediating role of task value. *Studies in Educational Evaluation*, 72, 101119.

Cheng, L., Wu, Y., & Liu, X. (2015). Chinese university students' perceptions of assessment tasks and classroom assessment environment. *Language Testing in Asia*, 5, 1-17.

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools*, 41(5), 535-548.

Colthorpe, K., Gray, H., Ainscough, L., & Ernst, H. (2021). Drivers for authenticity: Student approaches and responses to an authentic assessment task. *Assessment & Evaluation in Higher Education*, 46(7), 995-1007.

Dent, A. L., & Koenka, A. C. (2015). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, 28(4), 425-474.

Dorman, J. P., & Knightley, W. M. (2006). Development and validation of an instrument to assess secondary school students' perceptions of assessment tasks. *Educational Studies*, 32, 47-58.

Jin, Y., & Sun, H. (2020). Research on foreign language classroom assessment (2007-2018): Review and prospects. *Journal of Northeast Normal University (Philosophy and Social Sciences Edition)*, (5), 166-173.

Larenas, C. D., Boero, N. A., Rodríguez, B. R., & Sánchez, I. R. (2021). English language assessment: unfolding school students' and parents' views. *Educação e Pesquisa*, 47, e226529.

Lizzio, A., & Wilson, K. (2013). First-year students' appraisal of assessment tasks: implications for efficacy, engagement and performance. *Assessment & Evaluation in Higher Education*, 38(4), 389-406.

McMillan, J. H., & Workman, D. J. (1998). Classroom Assessment and Grading Practices: A Review of the. *Psychology*, 84, 261-271.

Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory Into Practice*, 44, 11 – 18.

Van Dinther, M., Dochy, F., Segers, M., & Braeken, J. (2014). Student perceptions of assessment and student self-efficacy in competence-based education. *Educational Studies*, 40(3), 330-351.

Velayutham, S., & Aldridge, J. M. (2013). Influence of psychosocial classroom environment on students' motivation and self-regulation in science learning: A structural equation modeling approach. *Research in Science Education*, 43(5), 507-527.

Zhang, W., & Li, Y. (2023). Development and validation of a questionnaire to assess classroom assessment from the self-regulated learning perspective. *Oxford Review of Education*, 49(6), 781-799.

e-ISBN: 978-629-98082-2-0

P005

Evaluating the Role of Local Large Language Models in Open-Ended Writing Assessment: Innovations, Validity, and Ethical Considerations in a Canadian University Context

Johanathan Woodworth

Faculty of Education, Mount Saint Vincent University, Canada.
(E-mail: johan.woodworth@msvu.ca)

Abstract

This study investigates the viability of using a locally deployed large language model (LLM), Qwen/qwen3-30b-a3b, for formative assessment of graduate-level writing in a Canadian university from a larger study that examined 4 different local models. Motivated by the underrepresentation of humanities-based assessment in generative AI research and growing concerns about data ethics in commercial platforms, the study deployed a 30-billion parameter model on university-secured infrastructure to assess thirty-five analytic assignments. Using a five-criteria rubric, both human raters and the LLM scored student work. While overall human scoring showed high reliability, the model's performance diverged significantly in categories demanding theoretical interpretation and ethical reasoning. In particular, the LLM failed to demonstrate meaningful score variance in Technology Analysis and underperformed in categories such as Ethical Considerations. Although its feedback was rubric-aligned and coherent, it lacked personalization and pedagogical depth. The study also incorporated autoethnographic reflection, revealing tensions around instructor identity, epistemic trust, and algorithmic mediation. Results underscore the limitations of mid-scale locally hosted LLMs for summative assessment and highlight the continued importance of teacher agency and hybrid feedback models. Ethical concerns related to algorithmic opacity, equity, and feedback decontextualization persist even under local hosting. A context-sensitive, pedagogically grounded approach remains essential for responsible AI integration in academic writing assessment.

Keywords: Local Large Language Models, Educational Assessment, Writing Feedback, Academic Integrity, Data Ethics

1. Introduction

The integration of generative AI into educational assessment represents a paradigmatic shift in how institutions approach the evaluation of student work, particularly in disciplines that rely

e-ISBN: 978-629-98082-2-0

on subjective, open-ended responses. This study addresses two persistent gaps in the literature: the underrepresentation of the humanities in AI-enabled assessment research and the ethical risks of commercial LLM platforms in educational contexts. By deploying LLMs locally on university-secured infrastructure, this research ensures data governance, reduces privacy risks, and explores broader applicability beyond STEM or short-answer domains.

This investigation is framed by a growing debate over AI in education. Proponents argue that engagement with AI supports digital literacy and fosters academic integrity (Chan, 2023b). Critics, however, warn that such integration may impair writing skills, critical thinking, and equitable learning outcomes (Warschauer et al., 2023; Crompton & Burke, 2023). These concerns are particularly salient in the domain of academic writing instruction, where process-oriented feedback and instructor-student dialogue have long been emphasized as essential to learning (Woodworth, 2023a; Woodworth, 2023b). While AI-generated feedback may offer efficiency and consistency, its capacity to support iterative writing development remains in question.

Further, prior work has examined the use of automated writing evaluation (AWE) systems in language classrooms and cautioned that over-reliance on algorithmic feedback may marginalize the pedagogical value of teacher-mediated intervention (Woodworth & Barkaoui, 2020). These critiques underscore the importance of preserving teacher agency in AI-mediated environments and inform the present study's emphasis on formative rather than summative applications.

The ethical and institutional implications of AI use in higher education demand more comprehensive governance frameworks. While international models such as UNESCO's tripartite framework offer foundational guidance (UNESCO, 2021), adapting these frameworks to local contexts requires critical attention to equity, autonomy, and infrastructure (Woodworth & Ballantyne, 2025). In response to the unpredictable and emergent nature of classroom interactions, this study draws on complexity theory as a theoretical anchor. Complexity theory emphasizes how learning environments are dynamic, non-linear, and shaped by interactions among multiple variables. Within the context of AI-mediated assessment, this perspective foregrounds the need for adaptive systems that remain sensitive to learner diversity, contextual factors, and the recursive nature of formative feedback. In addition, given the inherent unpredictability of educational interactions and learner diversity, complexity theory offers a valuable lens for understanding how language assessment practices, especially those involving generative AI, must remain adaptive, dialogic, and context-sensitive (Woodworth, 2025). This study investigates the performance of a high-parameter local language model (Qwen/qwen3-30b-a3b), deployed within a university-secured infrastructure for writing assessment. While the

e-ISBN: 978-629-98082-2-0

broader project includes other models, this paper focuses exclusively on Qwen due to its relative performance and relevance.

2. Methods

This mixed-methods study combined autoethnographic reflection with empirical analysis to evaluate AI-enabled writing assessment (Ellis & Bochner, 2000). Thirty-five graduate-level assignments were assessed using an analytic rubric by both human raters and a locally deployed large language model (LLM). The rubric included five criteria: Technology Analysis, Application of Theoretical Framework, Ethical Considerations, Integration of Technology and Framework, and Research and Writing Quality. Each was rated on a 5-point scale (with half-point increments), then normalized to a 20-point scale for analytical consistency. Assignments required students to critically examine an emerging educational technology and its classroom integration. All were scored independently by human raters and the LLM using the same rubric. Human rater consistency was assessed using the Intraclass Correlation Coefficient (ICC) via a two-way mixed-effects model (absolute agreement, single measures). ICC values ranged from 0.598 (Technology Analysis) to 0.818 (Research and Writing Quality). Theoretical Framework (0.713), Ethical Considerations (0.735), and Integration (0.725) showed good reliability. The composite Total Score yielded excellent agreement (ICC = 0.927), supporting the use of averaged human scores for comparison with AI outputs.

The LLM ran on university-secured Apple M3 Max hardware (16-core CPU, 64 GB RAM) via LM Studio. All data were encrypted and processed locally to ensure privacy compliance. Model selection prioritized pedagogical alignment, macOS compatibility, open-weight licensing, and suitability for both statistical and qualitative analysis. GGUF and MLX compatibility enabled quantized performance and native hardware integration. Qwen/qwen3-30b-a3b (Qwen3 MoE, 4-bit) was deployed for its parameter count, instruction-following ability, and infrastructure compatibility. An instruction-following prompt was used due to its efficiency, low latency, and alignment with rubric-based tasks. Alternatives such as Chain-of-Thought, Long Context, or Agent Workflows were excluded due to unnecessary processing demands or complexity for the scoring task (Murthy et al., 2024; Lou et al., 2024).

Quantitative analysis included paired-samples t-tests, Wilcoxon signed-rank tests, and correlation analysis to compare average human and AI scores. Qualitative analysis used thematic coding on ~200 words of AI-generated feedback per assignment. Model-human agreement was examined alongside autoethnographic insights into deployment challenges and broader cultural tensions. Iterative prompt refinement balanced specificity with hallucination mitigation. This initial phase of the larger research project offers a comprehensive appraisal of

e-ISBN: 978-629-98082-2-0

feedback quality, scoring validity, and ethical implications. By combining statistical rigor, qualitative depth, and reflexive inquiry, the study provides a robust evaluation of the pedagogical potential of locally hosted LLMs in academic assessment.

3. Results and Discussion

3.1 Grading Performance and Validity

AI-generated scores were compared to human ratings across five analytic rubric categories and the total score. Using the Qwen/qwen3-30b-a3b model, statistically significant differences emerged in three categories. For Application of Theoretical Framework, human scores were on average 2.4 points higher, $t(34) = 3.75$, $p < .001$. For Ethical Considerations, the gap widened to 3.0 points, $t(34) = 9.45$, $p < .001$. Similarly, in Integration of Technology and Framework, human scores surpassed AI scores by 1.9 points, $t(34) = 4.86$, $p < .001$. These results were supported by Wilcoxon signed-rank tests ($p < .01$ in all three cases). In contrast, no significant differences were found in Technology Analysis or Research and Writing Quality. For the latter, the mean difference was 0.2 points, $t(34) = -0.56$, $p = .577$. The uniformity in Technology Analysis, however, raises concern: the model assigned a flat score of 16.0 across all 35 assignments, yielding $t(34) = 0.00$, $p = 1.00$. This lack of score variation indicates a fundamental limitation in the model's evaluative sensitivity. Correlation analyses confirmed this misalignment. Pearson coefficients across categories ranged from −.26 to .24, none statistically significant. For Technology Analysis, correlation was undefined due to invariant scoring. Spearman rank-order correlations were similarly weak, with only Ethical Considerations showing a significant negative correlation ($\rho = -.39$, $p = .019$).

These results indicate the model's failure to approximate human rank-order judgments, especially in categories requiring conceptual or ethical reasoning. It systematically under-ranked higher-quality work, particularly in Ethical Considerations. While it matched human ratings more closely in surface-level domains like Research and Writing Quality, it performed poorly in dimensions requiring theoretical or interpretive nuance. Its failure to differentiate scores in Technology Analysis further underscores this insensitivity. Although the model may provide utility for low-stakes formative feedback, its compressed scoring range and lack of responsiveness to performance variation undermine its validity for summative use. These findings contrast with studies showing closer alignment between human and AI scoring under controlled conditions (Chiang & Lee, 2023; Mendonça et al., 2025), possibly due to the architectural limits of sub-32B models like Qwen. Notably, the model exhibited a "proportional bias," underrepresenting both high- and low-performing work, a known issue in automated writing assessment (Wetzler et al., 2025). While such models may enhance efficiency in formative contexts, their inability to capture originality, interpretive depth, or theoretical rigor

e-ISBN: 978-629-98082-2-0

continues to limit their appropriateness for high-stakes evaluation (Hao et al., 2024). These findings reaffirm that current mid-scale LLMs remain inadequate where nuanced performance differentiation is essential.

3.2 Feedback Quality and Pedagogical Value

The AI-generated feedback from the Qwen model displayed limitations. While the model generated coherent and rubric-aligned comments, it still lacked the personal tone, contextualization, and dialogic engagement that are hallmarks of effective formative feedback in writing-intensive disciplines (Seßler et al., 2025; Woodworth, 2023a). The feedback from the model tended to be generic, often reiterating rubric points without tailoring comments to individual student strengths or areas for growth or were generic in nature across students with an overly positive tone.

These findings reinforce the necessity of hybrid approaches, where AI-generated comments serve as a supplement to, rather than a replacement for, instructor feedback. Human mediation is essential to preserve pedagogical intentionality, ensure relevance, and maintain student engagement and trust (Woodworth, 2023b; Bibi et al., 2024).

3.3 Ethical and Epistemological Considerations

The autoethnographic lens of the study revealed persistent tensions regarding trust, instructor identity, and epistemic authority in AI-mediated assessment. Throughout the study, the researcher was persistently uneasy about how the AI would score assignments and whether its feedback was genuinely effective. The researcher observed that the AI's feedback was almost always positive, sometimes to the point of being bland or uncritical, which led to questions about its usefulness for student growth. This overly positive tone appeared disconnected from the realities of student work and risked undermining the credibility of the assessment process. As a teacher, the researcher experienced a shift in professional identity, no longer serving as the sole authority on student writing but rather acting as a mediator between human judgment and algorithmic output. This raised concerns about the teacher's role, trust in the process, and the limits of delegating such an important part of teaching to AI. These reflections highlight the ongoing need for teacher agency, careful mediation of AI feedback, and a cautious approach to integrating AI into assessment practice. These affective and epistemological dimensions are increasingly recognized as central to the adoption of AI in education (Kizilcec, 2024; Naseri & Abdullah, 2024).

Further, ethical risks related to training data provenance, opacity in model fine-tuning, and the decontextualization of feedback align with UNESCO's call for AI deployments that are "inclusive, transparent, and contextually grounded" (UNESCO, 2021). The use of locally

e-ISBN: 978-629-98082-2-0

hosted LLMs provided enhanced infrastructural control and data privacy, mitigating some risks associated with commercial platforms (Hanke et al., 2024). However, institutional adoption of such tools does not confer ethical immunity; robust, context-specific governance frameworks remain essential to address issues of bias, accountability, and equitable access (Woodworth & Ballantyne, 2025; Baker & Hawn, 2021).

Collectively, these results suggest that while local LLMs hold promise for augmenting formative feedback processes, The Qwen model's current limitations, particularly in personalization, interpretive depth, and ethical transparency, preclude their use as standalone tools for high-stakes, summative assessment. The findings support a cautious, context-sensitive approach to AI integration in educational assessment, emphasizing the continued centrality of human judgment and pedagogical expertise (Woodworth, 2023a, 2023b).

4. Conclusion

This study offers empirical insights into the pedagogical viability and ethical complexities of using locally run large language models (LLMs) for formative writing assessment in a Canadian university. The findings from the Qwen model diverge from earlier studies suggesting LLMs can match human grading in controlled conditions (Chiang & Lee, 2023; Mendonça et al., 2025). While surface-level agreement emerged in low-variance domains like mechanics, the locally hosted model showed clear limits in assessing interpretive nuance, theoretical reasoning, and ethical reflection. These may stem from the Qwen model's smaller parameter count relative to commercial systems like OpenAI's GPT-4. Larger models gain from scale, reinforcement learning, and continuous optimization, which are factors likely supporting superior assessment outcomes.

The Qwen model also showed proportional bias, consistently underrepresenting both high- and low-performing submissions. This aligns with enduring concerns in automated scoring research and was particularly clear in uniform scores for areas such as Technology Analysis. The model's inability to capture performance variation undermines its reliability for nuanced, rubric-based assessment. These findings underscore the need for hybrid approaches, where AI-generated scores and feedback serve as instructional aids rather than substitutes for human judgment. When used critically, such tools can improve efficiency while preserving the intentionality, responsiveness, and trust vital to formative assessment (Seßler et al., 2025; Jansen et al., 2024; Bibi et al., 2024; Woodworth, 2023b).

While the Qwen/qwen3-30b-a3b model produced rubric-aligned feedback in lower-order categories, its performance in higher-order domains remained erratic. Shortcomings in

e-ISBN: 978-629-98082-2-0

capturing originality, interpretive depth, and score variance persist. Though based on one model and dataset, these limitations may shift as further empirical evidence emerges. Still, the model's tendency to compress score distributions and underrepresent outliers remains a challenge for deploying GenAI in high-stakes summative contexts (Hao et al., 2024; Wetzler et al., 2025).

Ethical concerns, such as algorithmic bias, data opacity, and decontextualized feedback, remain relevant even with institutionally hosted LLMs. While local deployment offers more privacy and control than commercial platforms (Hanke et al., 2024; UNESCO, 2021), it does not remove the need for strong governance. Institutional policies must continue to emphasize equity, accountability, and pedagogical intentionality in AI integration (Woodworth & Ballantyne, 2025; Baker & Hawn, 2021).

The study's reflexive autoethnographic lens provided insight into the affective and epistemic dimensions of AI-mediated assessment, highlighting how trust, instructor identity, and professional agency are negotiated when integrating algorithmic tools. These findings exposed not only technical and pedagogical issues, but also emotional and ethical tensions shaping instructor and student experiences.

This autoethnographic view also illuminated ongoing tensions around trust, identity, and epistemic authority in AI assessment. These dynamics emphasize the need for governance frameworks that prioritize transparency, professional agency, and student-centered learning (Kizilcec, 2024; Naseri & Abdullah, 2024). Complexity theory reinforces that AI integration must be seen as an evolving process. In dynamic learning ecosystems where feedback loops and student-teacher interactions shape outcomes, rigid or universal AI solutions risk oversimplifying educational complexity. Effective assessment must remain context-sensitive, adaptive, and grounded in iterative practice. These principles should guide both LLM development and instructional design.

Several limitations warrant note: the sample was limited to thirty-five graduate assignments from a single institution and discipline; the focus was on analytic, rubric-based tasks rather than creative writing; and only one model and prompting strategy were used. Future research should expand to multiple genres, institutional contexts, and LLMs, as well as explore alternative prompting. Ongoing challenges in personalizing feedback, managing subjectivity, and capturing interpretive nuance also mark critical areas for refinement.

Moving forward, institutions should invest in transparency, faculty development, and infrastructure to harness the pedagogical value of AI while mitigating associated risks (Chan, 2023a; Hanke et al., 2024). Future research should investigate co-designed feedback

e-ISBN: 978-629-98082-2-0

mechanisms that integrate AI with human expertise, explore culturally responsive and multilingual models, and refine AI-assisted rubrics to better capture complexity and diversity in academic writing (Woodworth, 2025). Additionally, alternative prompting strategies, such as chain-of-thought or context-aware techniques, may enhance precision, score variability, and feedback personalization beyond the instruction-following approach employed in this study.

Ultimately, situating AI adoption within a pedagogical and ethical framework that centers teacher agency, learner diversity, and institutional transparency will enable educational institutions to harness the benefits of generative AI while mitigating its risks. A context-sensitive, adaptive approach will support more inclusive, trustworthy, and effective assessment practices (UNESCO, 2021; Baker & Hawn, 2021).

Acknowledgement

This research was supported by an internal research grant from Mount Saint Vincent University. The author thanks the Faculty of Education and the Research Office for their support in providing infrastructure, funding, and institutional guidance throughout the study.

References

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bibi, A., Yamin, S., Natividad, L. R., Rafique, T., Akhter, N., Fernandez, S. F., & Samad, A. (2024). Navigating the ethical landscape: AI integration in education. *Educational Administration: Theory and Practice*, *30*(6), 1579-1585.

Chan, C. K. Y. (2023a). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*(1), 38. https://doi.org/10.1186/s41239-023-00408-3

Chan, C. K. Y. (2023b). Is AI changing the rules of academic misconduct? An in-depth look at students' perceptions of 'AI-giarism'. *arXiv*. https://arxiv.org/abs/2306.03358

Chiang, C. H., & Lee, H. Y. (2023). Can large language models be an alternative to human evaluations?. *arXiv preprint arXiv:2305.01937*.

Crompton, H., & Burke, D. (2023). Academic integrity and artificial intelligence: Challenges and responses. *International Journal of Educational Integrity, 19*, 7. https://doi.org/10.1007/s40979-023-00128-9

e-ISBN: 978-629-98082-2-0

Ellis, C., & Bochner, A. P. (2000). Autoethnography, personal narrative, reflexivity: Researcher as subject. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 733–768). Sage Publications.

Hanke, V., Blanchard, T., Boenisch, F., Olatunji, I. E., Backes, M., & Dziedzic, A. (2024). Open LLMs are necessary for current private adaptations and outperform their closed alternatives. *arXiv*. https://arxiv.org/abs/2411.05818

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI. *Educational Measurement: Issues and Practice*, *43*(2), 16–29. https://doi.org/10.1111/emip.12602

Kizilcec, R. F. (2024). To Advance AI Use in Education, Focus on Understanding Educators. *International Journal of Artificial Intelligence in Education*, 1–8. https://doi.org/10.1007/s40593-023-00351-4

Lou, R., Zhang, K., & Yin, W. (2024). Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, *50*(3), 1053-1095.

Mendonça, P. C., Quintal, F., & Mendonça, F. (2025). Evaluating LLMs for automated scoring in formative assessments. *Applied Sciences*, *15*(5), 2787.

Murthy, R., Kumar, P., Venkateswaran, P., & Contractor, D. (2024). Evaluating the Instruction-following Abilities of Language Models using Knowledge Tasks. *arXiv preprint arXiv:2410.12972*.

Naseri, R. N. N., & Abdullah, M. S. (2024). Understanding AI technology adoption in educational settings: A review of theoretical frameworks and their applications. *Information Management and Business Review*, *16*(3), 174-181.

Seßler, K., Bewersdorff, A., Nerdel, C., & Kasneci, E. (2025). Towards Adaptive Feedback with AI: Comparing the Feedback Quality of LLMs and Teachers on Experimentation Protocols. *arXiv*, 2502.12842v1. http://arxiv.org/abs/2502.12842v1

Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Empirische Arbeit: Comparing Generative AI and Expert Feedback to Students' Writing: Insights from Student Teachers. *Psychologie in Erziehung und Unterricht*, *71*(2), 11. https://doi.org/10.2378/peu2024.art08d

United Nations Educational, Scientific and Cultural Organization. (2021). *AI and education: Guidance for policy-makers*. https://unesdoc.unesco.org/ark:/48223/pf0000376709

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for second language writers. *SSRN*. https://doi.org/10.2139/ssrn.4404380

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2025). Grading the Graders: Comparing Generative AI and Human

e-ISBN: 978-629-98082-2-0

Assessment in Essay Evaluation. *Teaching of Psychology*, *52*(3), 298–304. https://doi.org/10.1177/00986283241282696

Woodworth, J., & Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal,* 37(2), 234-247.

Woodworth, J. (2023a). Product to process: The efficacy of hybrid feedback in academic writing classrooms for fostering process-oriented writing. In S. Chong & H. Reinders (Eds.), *Innovation in learning-oriented language assessment* (pp. 295-310). Springer International Publishing.

Woodworth, J. (2023b). The use of hybrid corrective feedback in the foreign language classroom and its implications for teacher education: A case study with Chinese learners in an EAP setting. In P. Hohaus & J. F. Heeren (Eds.), *The Future of Teacher Education* (pp. 200-227). Brill.

Woodworth, J. (2025, June). Navigating complexity in language assessment: Adapting to learner needs through dynamic approaches [Conference presentation]. Symposium: Complexity theory and adaptive approaches to language assessment (Chair: J. Woodworth; Discussant: M. Hunte), Association canadienne de linguistique appliquée / Canadian Association of Applied Linguistics (ACLA/CAAL) Annual Conference, Toronto, Canada.

Woodworth, J., & Ballantyne, E. (2025, August). Adapting the AI Ecological Education Policy Framework to the Canadian context [Conference presentation]. IEEE Services Workshop on Societal and Ethical Governance of AI (SEGA 2025), Calgary, Canada.

e-ISBN: 978-629-98082-2-0

P029

Exploring Rater Effects in Automated Assessment of EFL Learners' Paraphrasing Skills with NLP Metrics and Customized GPT

***Minkyung Kim[1]**

[1]English Language and Literature, Seoul National University, South Korea.
(E-mail: [1]minkyung.kim0322@snu.ac.kr)

Abstract

Paraphrasing is a critical component of academic writing, enabling EFL learners to demonstrate lexical variation, syntactic flexibility, and semantic control. This study investigates the validity and reliability of automated paraphrasing assessment by comparing three scoring systems: (1) human raters (Hm), (2) a custom GPT-based model (MyGPT), and (3) NLP metric-based scoring. Using a dataset of 1,000 sentence-level paraphrases produced by 100 Korean EFL learners, each response was evaluated using an analytic rubric comprising three dimensions: Syntactic Change, Word Change, and Semantic Similarity. Many-Facet Rasch Measurement (MFRM) was used to examine the psychometric properties of each rating system. The analysis revealed that NLP metrics were the most severe, followed by human raters, with MyGPT scoring most leniently. Among the rubric criteria, Syntactic Change was the most challenging for learners. Pearson correlations across rating systems showed moderate to high reliability, with particularly strong alignment between MyGPT and human ratings. The findings suggest that GPT-based scoring models can generate rubric-aligned and reliable evaluations, often with a more lenient severity profile than both human raters and traditional NLP metrics. While this leniency may reflect GPT's flexible interpretation of rubric criteria, careful prompt calibration is needed to maintain scoring consistency. These results highlight the importance of psychometric validation in automated assessment and offer pedagogical implications for the responsible integration of AI-based evaluation in EFL writing instruction.

Keywords: paraphrase assessment, Many-Facet Rasch Measurement, ChatGPT, natural language processing, automatic scoring

1. Introduction

Writing assessment has become a critical area of inquiry in second language acquisition (SLA) research, as it directly evaluates learners' ability to produce language in meaningful and contextually appropriate ways. One persistent challenge in writing assessment is ensuring rater

e-ISBN: 978-629-98082-2-0

reliability, particularly in the face of rater bias—systematic tendencies of raters to be overly severe or lenient regardless of the actual quality of student writing (Engelhard, 1994; McNamara, 1996). Such bias can compromise the fairness and validity of assessment results and remains a central concern in both large-scale and classroom-based evaluations (Lynch &McNamara, 1998; Kondo-Brown, 2002).

This issue is also salient in paraphrasing assessment, a subdomain of writing that is gaining increasing pedagogical importance. Paraphrasing tasks require learners to restate source content using different vocabulary and syntactic structures while preserving the original meaning (Keck, 2006; Shi, 2004). In EFL contexts, paraphrasing is not only a tool for developing lexical and grammatical competence but also a safeguard against plagiarism, which is particularly relevant in academic settings (Campbell, 1990). Despite its instructional value, paraphrasing remains under-assessed, and the absence of widely accepted scoring rubrics often leads to inconsistencies in rating (Liu & Lin, 2022).

At the same time, automated scoring systems are becoming more prevalent in writing assessment, especially with the rapid development of generative AI technologies. Large language models (LLMs) like ChatGPT are now being used to provide formative feedback and even summative evaluations. While these systems offer the promise of scalable, cost-effective, and consistent scoring, they also raise new questions: Do AI-based scorers exhibit rater bias, and if so, how does it compare to that of human raters or traditional NLP-based scoring systems? Understanding the rater characteristics—such as severity, leniency, and consistency—of different scoring methods is essential for determining whether these tools are valid and equitable alternatives to human judgment.

The current study addresses this critical gap by comparing three scoring approaches in paraphrasing assessment: (1) traditional human ratings, (2) scores generated by a custom ChatGPT-based model (MyGPT), and (3) scores calculated using natural language processing (NLP) metrics. Employing Many-Facet Rasch Measurement (MFRM; Linacre, 1989) this study investigates how each scoring system behaves as a rater by examining patterns of severity, rater fit, dimensional consistency, and overall reliability. MFRM allows for the analysis of multiple facets simultaneously, including item difficulty, participant ability, rater severity, and scoring criteria (Engelhard, 1992; Eckes, 2005; Kim, 2009; Shin, 2010; Lim, 2011). By conceptualizing AI and NLP systems as evaluators in their own right, the study asks whether these systems fulfill the psychometric standards expected of human raters.

Beyond technical analysis, the study offers practical implications for EFL educators and assessment designers. As AI-powered evaluation tools are increasingly integrated into

e-ISBN: 978-629-98082-2-0

classrooms and testing environments, it becomes crucial to establish guidelines for their use, limitations, and calibration. Understanding how MyGPT and NLP-based scores align or diverge from human judgments allows for more informed decisions about adopting such systems, especially in high-stakes contexts like writing instruction and proficiency testing.

In doing so, this research contributes to the evolving discussion on the role of AI in education, shedding light on its potential not only as a tool for language generation but also as a rater with identifiable characteristics and measurable reliability.

2. Methods

This study adopts a mixed-methods design to examine and compare the scoring behavior of three distinct evaluation systems used in paraphrasing assessment: human raters, a custom GPT-based model (MyGPT), and NLP metric-based scoring. A total of 1,000 paraphrased responses were collected from 100 Korean EFL learners, and each response was evaluated across four analytic dimensions: syntactic complexity, lexical diversity, semantic similarity, and mechanical accuracy. Employing Many-Facet Rasch Measurement (MFRM), the study analyzed how each scoring system functioned as a rater by investigating severity, rater fit, dimensional consistency, and inter-rater reliability. The analysis aimed to identify patterns in scoring behavior and to evaluate the psychometric validity of automated scoring methods in comparison to human judgment.

2.1 Participants

A total of 20 Korean EFL learners participated in this study and completed the paraphrasing test developed for research purposes. All participants were university students (both undergraduate and graduate), with ages ranging from early twenties to mid-thirties (M = 24.52).

2.2 Materials

The paraphrasing task developed for this study consisted of five items, and each item included two sentences to be paraphrased. Participants were instructed to produce one paraphrased version for each sentence, resulting in 10 paraphrasing tasks per participant. With 100 participants completing the task, the dataset comprised a total of 1,000 paraphrased sentences.

2.3 Scoring Rubric

An analytic scoring rubric was used to assess each paraphrase across three dimensions: Syntactic Change (SC), Word Change (WC), and Semantic Similarity (SS). Each dimension was rated on a 6-point scale from 0 to 5. The analytic rubric was selected for its ability to provide more detailed and diagnostic insights into learner performance.

e-ISBN: 978-629-98082-2-0

2.4 Rating Systems

Three different evaluation systems were used to score the paraphrased responses:

Human Raters: Two trained raters with expertise in English language assessment independently rated all responses using the analytic rubric. Inter-rater reliability reached a Pearson correlation coefficient of .81. Final scores were determined through consensus.

MyGPT: A custom-prompted version of ChatGPT was used to assign scores to each paraphrase based on the same analytic rubric. Prompt engineering ensured alignment with the original scoring dimensions.

NLP Metric-Based Scoring: Three NLP metrics were employed to approximate the analytic rubric dimensions. Metric outputs were normalized to match the 0–5 rubric scale.
All systems reached high inter-rater reliability in separate analysis.

2.5 Analysis

MFRM analysis was used to analyze rater characteristics across the three scoring systems, making it suitable for evaluating the comparability of human and automated ratings. This analysis included three facets: (1) examinees (meaning participants), (2) rating systems (Human, MyGPT, NLP), and (3) scoring criteria (Syntactic Change, Word Change, Semantic Similarity). In addition, to investigate inter-rater reliability, Pearson correlation coefficients (PCCs) were calculated for each scoring criterion.

3. Results and Discussion

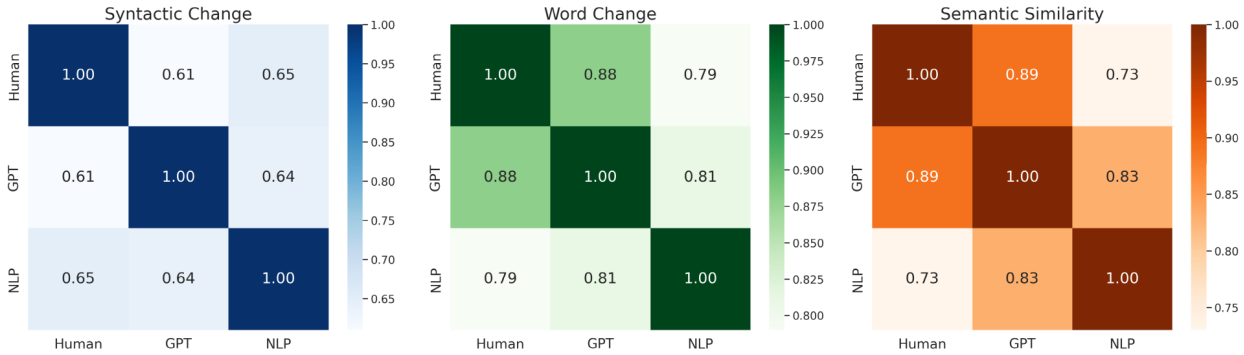
This study examined the scoring behaviors of three evaluation systems—Human Raters, a custom GPT-based model (MyGPT), and NLP Metric-based scoring. Inter-rater reliability among different rating systems was investigated with PCC values. Figure 1 presents the Pearson correlation coefficients among the three rating systems—Human Rater, GPT, and NLP Metrics—across the three analytic criteria. These coefficients reflect the degree of linear association between scoring systems, offering insight into their alignment across three scoring dimensions. All of the values were statistically significant (p > .01).

For Syntactic Change, the Pearson correlation between Human and GPT was relatively modest (r = .61), as was the correlation between Human and NLP Metrics (r = .65), and GPT and NLP (r = .64). These moderate correlations suggest that syntactic modifications are interpreted differently by each system, potentially due to differences in how structural transformation is conceptualized or operationalized.

e-ISBN: 978-629-98082-2-0

In contrast, Word Change exhibited higher correlations across all pairings.

Figure 1: Inter-Rater Reliability Among Human, GPT, and NLP Scoring Systems by Analytic Criterion

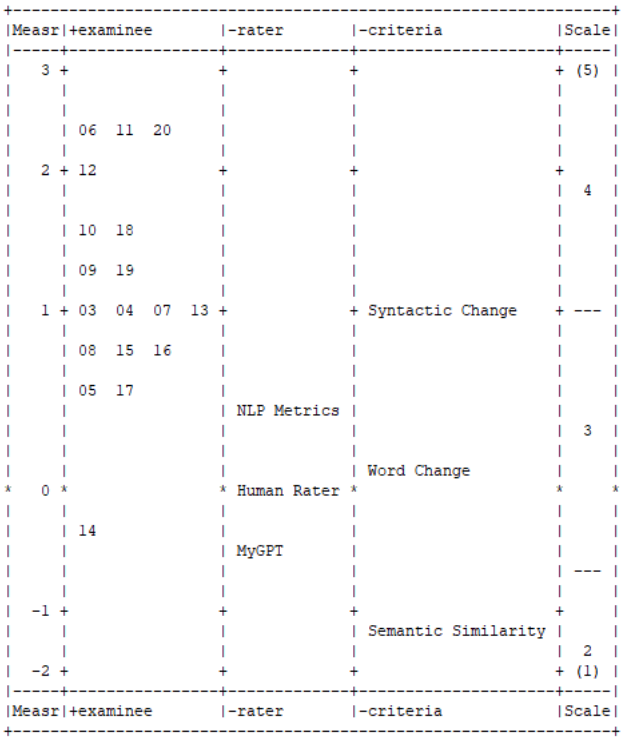


The correlation between Human and GPT reached .88, indicating a strong linear relationship. The correlation between Human and NLP was .79, and between GPT and NLP was .81. These results demonstrate that lexical variation is evaluated more consistently across systems, possibly due to its relatively surface-level nature and detectability by both rule-based and neural models.

The highest correlations were found in Semantic Similarity, with Human–GPT at .89, GPT–NLP at .83, and Human–NLP at .73. This suggests that GPT's semantic assessment is highly aligned with human judgment, likely due to its capacity to capture contextual meaning through deep language modeling. In sum, Pearson correlations reveal that GPT scores show higher agreement with human ratings than traditional NLP metrics do, particularly in Word Change and Semantic Similarity. However, Syntactic Change remains a challenging dimension, with only moderate correlations observed across all rating pairs.

Figure 2: All Vertical Rulers for Sample 20 Participants

e-ISBN: 978-629-98082-2-0



The scoring behaviors of three evaluation systems—Human Raters, a custom GPT-based model (MyGPT), and NLP Metric-based scoring— were investigated through a Many-Facet Rasch Measurement (MFRM) analysis. Figure 2 presents a vertical ruler that aligns the estimated measures of three key facets—examinees, raters, and scoring criteria—on a common logit scale, facilitating direct comparisons of their relative severity or proficiency. Regarding rater severity, NLP Metrics were found to be the most severe, appearing at the top of the scale. Human raters occupied a mid-level position, while the MyGPT was located lower on the scale, indicating a more lenient scoring tendency.

In terms of criteria difficulty, Syntactic Change was positioned higher on the scale than the other two, indicating that it was the most challenging aspect for examinees. In contrast, Word Change and Semantic Similarity were placed lower, suggesting they were relatively easier constructs or were assessed more leniently across raters.

The MFRM analysis reveals important differences in scoring behavior across the three evaluation systems. The NLP-based scoring system showed the highest severity, followed by MyGPT, while human raters were the most lenient. This finding highlights the need for careful calibration when integrating automated scoring into assessment frameworks, especially to avoid penalizing learners due to overly strict algorithms.

In terms of scoring criteria, syntactic change was consistently rated as the most difficult aspect of paraphrasing, suggesting that structural transformation poses a greater challenge for

e-ISBN: 978-629-98082-2-0

learners compared to lexical or semantic changes. This pattern holds across all raters, implying that syntactic complexity remains a core area of difficulty regardless of the evaluator.

Finally, while the sample of examinees exhibited a range of abilities, the clustering of many participants in the mid-logit range suggests a relatively homogeneous proficiency level in the current sample. Future studies may consider a broader range of learner profiles to further investigate rating stability across ability groups.

4. Conclusion

This study investigated the scoring characteristics of Human Raters, MyGPT, and NLP Metrics in evaluating English paraphrasing performance. The Many-Facet Rasch Measurement (MFRM) and correlation analyses revealed a clear severity hierarchy: NLP-based scoring was the most stringent, followed by human raters, with the customized GPT being the most lenient. Syntactic Change emerged as the most difficult criterion, highlighting a persistent challenge for EFL learners in restructuring sentence form while maintaining meaning.

These findings have several educational implications. First, the differential severity among scoring systems emphasizes the importance of selecting or combining appropriate evaluators depending on the assessment purpose—whether diagnostic, summative, or formative. Second, the consistently low performance on syntactic change tasks suggests a need for explicit instruction and practice in syntactic transformation, not just lexical substitution. Third, while NLP-based systems offer efficiency and scalability, their harshness may discourage learners if used without supportive feedback or scaffolding. MyGPT, on the other hand, shows promise as a more balanced automated rater that approximates human judgment.

In sum, AI-assisted scoring systems like MyGPT can serve as effective tools for paraphrasing assessment, particularly when integrated with educational strategies that foster syntactic awareness and provide learner-friendly feedback. Future work should continue refining such systems to align more closely with both linguistic validity and pedagogical goals.

Acknowledgement

This research was conducted without external funding. I express huge gratitude to the AALA 2025 conference organizers for providing the opportunity to present and discuss this study.

e-ISBN: 978-629-98082-2-0

References

Campbell, C. (1990). Writing with others' words: Using background reading text in academic compositions. In B. Kroll (Ed.), Second language writing: Research insights for the classroom (pp. 211–230). Cambridge, UK: Cambridge University Press.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. Language Assessment Quarterly, 2(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5(3), 171–191. https://doi.org/10.1207/s15324818ame0503_2

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. Journal of Educational Measurement, 31(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. Journal of Second Language Writing, 15(4), 261–278. https://doi.org/10.1016/j.jslw.2006.09.001

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. Language Testing, 26(2), 187–217. https://doi.org/10.1177/0265532208101006

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing, 19(1), 3–31. https://doi.org/10.1191/0265532202lt217oa

Lim, H. (2011). A many-facet Rasch analysis of the validity of a college English writing assessment. English Language Teaching, 23(1), 115–137.

Linacre, J. M. (1997). Guidelines for rating scales. MESA Research Note #2. Retrieved September 10, 2011, from http://www.rasch.org

Liu, S., & Lin, D. (2022). Developing and validating an analytic rating scale for a paraphrase task. Assessing Writing, 53, 100646. https://doi.org/10.1016/j.asw.2022.100646

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. Language Testing, 15(2), 158–180. https://doi.org/10.1177/026553229801500202

McNamara, T. F. (1996). Measuring second language performance. Longman.

Shi, L. (2004). Textual borrowing in second-language writing. Written Communication, 21(2), 171–200. https://doi.org/10.1177/0741088303262846

Shin, Y. S. (2010). A FACETS analysis of rater characteristics and rater bias in measuring L2 writing performance. English Language & Literature Teaching, 16(1), 123–142.

e-ISBN: 978-629-98082-2-0

P033

Language Assessment Knowledge of English Language Teachers in Assessing Four Language Skills

Nway Htway Khin

Graduate School of Language and Linguistics, Sophia University, JAPAN.
nwayhtwaykhinnancy1996@eagle.sophia.ac.jp

Abstract

Assessment is a crucial aspect of language teaching and learning, helping teachers track student progress, refine teaching methods, and support learner development. However, while general assessment knowledge among teachers has been studied, less attention has been given to the specific knowledge required to assess different language skills: reading, writing, listening, and speaking. This study explores the Language Assessment Knowledge (LAK) of English language teachers, focusing on their ability to assess these four skills and examining whether factors like teaching experience, educational background, and professional training influence their assessment knowledge. Using a quantitative research approach, data were collected from 12 English language teachers through a 60-item questionnaire. The results showed that teachers had moderate LAK levels, and teachers were more confident in assessing reading and speaking, but found writing and listening assessments more challenging. The study also found no significant relationship between teachers' LAK levels and their experience, education, or how often they assessed a particular skill. This suggests that having more teaching experience or higher academic qualifications does not necessarily lead to better assessment knowledge. Therefore, the study recommends integrating language assessment training into teacher education programs and offering professional development workshops to improve teachers' ability to design and implement effective language assessments, particularly in writing and listening assessments. Future research should expand sample sizes and include qualitative methods such as interviews and classroom observations to better understand how teachers apply assessment knowledge in practice.

Keywords: language assessment knowledge, language skills assessment, teacher training, professional development, English language teaching.

e-ISBN: 978-629-98082-2-0

1. Introduction

Assessment plays a vital role in teaching and learning as it serves as a driving force to help teachers track students' progress, improve teaching methods, and support learners in identifying their areas for improvement to achieve their goals. Brown (2003) makes an important distinction between testing and assessment, emphasizing that these are not interchangeable concepts. While testing is a formal and summative process where learners know they are being tested, typically in a controlled environment, assessment is an ongoing process that happens continuously throughout the teaching and learning process. Unlike testing, the ongoing nature of assessments allows teachers to make adjustments and improvements in real-time, addressing students' individual needs and fostering their progress effectively.

Moreover, the distinction between testing and assessment becomes even more significant when considering their purposes. Clapham and Corson (1997) add to this distinction by explaining that testing is often used for large-scale purposes, such as standardized exams, to compare the performance of many learners. On the other hand, assessment focuses on understanding individual learners' challenges and offering meaningful feedback to support their needs. This difference highlights how assessment can support personal growth and development, rather than simply measuring performance.

In the context of language education, assessment has become crucial in supporting second language acquisition. According to Purpura (2016), language assessment involves systematically collecting information about learners' language abilities to make informed decisions about instruction. Through effective assessment, teachers can help students achieve proficiency and communicative competence. However, despite its significance, many teachers face challenges in designing and implementing effective assessment tools, particularly due to a lack of specialized knowledge in this area.

One of the critical challenges is the need for specialized knowledge, referred to as Language Assessment Knowledge (LAK). LAK includes the skill-specific knowledge required to design and implement assessments that address the unique demands of each language skill (Ölmezer-Öztürk, 2018). While the general assessment literacy of teachers has been widely studied, less attention has been given to the specific knowledge required for assessing learners' language skills, such as reading, writing, listening, and speaking. Ölmezer-Öztürk (2018) argues that skill-specific LAK is essential for teachers to develop assessments that address learners' needs and promote their progress. Without adequate language assessment knowledge, teachers may face difficulties in designing assessments that accurately measure learners' abilities or provide feedback for their improvement.

e-ISBN: 978-629-98082-2-0

Additionally, another pressing issue that requires more attention is the influence of demographic factors, such as teaching experience, educational background, and teaching qualification, on the teachers' assessment of knowledge. Mertler and Campbell (2005) and Fulcher (2012) argue that these factors can significantly impact a teacher's ability to design and implement assessments. While some studies have explored these factors, the relationships between them and teachers' ability to design effective assessments remain underexplored. Thus, addressing these gaps is essential for enhancing the quality of language assessments and ensuring that assessments meet learners' diverse needs in language learning.

To address these research gaps, this study focuses on investigating the Language Assessment Knowledge (LAK) of English language teachers, particularly their ability to assess the four core language skills: reading, writing, listening, and speaking. Additionally, it examines how demographic factors influence teachers' knowledge of language assessment. Hence, it is hoped that this study will be helpful for teachers in the language teaching context by contributing to improving language assessment practices and teacher training programs.

1.1 Abbreviations and Acronyms

LAK - Language Assessment Knowledge

LAL - Language Assessment Literacy

2. Methods

A quantitative research design was used to investigate the skill-based Language Assessment Knowledge (LAK) levels of English language teachers in assessing four language skills: reading, listening, writing, and speaking. The study focused on identifying English language teachers' knowledge of language assessment and examining potential relationships between their LAK levels and demographic factors such as teaching experience, educational background, and professional development. The participants consisted of 12 English language teachers from diverse professional backgrounds, including both native and non-native speakers. Half of the participants had 1–5 years of teaching experience, and the remaining half had 6–10 years. In terms of qualifications, eight held a bachelor's degree while four had a master's degree. Eight participants had specialized qualifications in ELT/TEFL/TESOL/CELTA, and four held general education degrees. The participants taught in various contexts: five in private institutions, one in a public institution, three as freelance or part-time teachers, and three were not currently teaching. Only three participants had attended language assessment-related professional development programs, while the majority had not.

Data were collected using a 60-item questionnaire developed by Ölmezer-Öztürk (2018). The

e-ISBN: 978-629-98082-2-0

instrument examined teachers' knowledge in assessing the four language skills, with 15 items for each skill: reading, listening, writing, and speaking. Participants were asked to respond to each item using a true/false/not given scale to indicate whether they had the knowledge or ability to perform specific assessment tasks. A printed version of the questionnaire was distributed in person to the participants, and informed consent was obtained from all participants.

Descriptive statistics (mean, standard deviation) were calculated for each skill. The midpoint for each skill was 7.5 out of 15, and for the total score, 30 out of 60. Point-biserial correlations were used to examine relationships between LAK levels and demographic variables, as well as between skill-specific LAK and frequency of assessment practices.

3. Results and Discussion

The overall mean LAK score was 28 out of 60, slightly below the midpoint of 30, indicating moderate but inconsistent proficiency. Reading (M = 8.58) and speaking (M = 8.00) scores were above the skill-specific midpoint, while listening (M = 6.17) and writing (M = 6.08) were below.

Table 1: Descriptive Statistics for Skill-Based LAK Levels of English Language Teachers Source: Author's analysis of questionnaire data

Skill	Total Score	Mean (M)	Standard Deviation (SD)
Reading	103	8.58	0.421
Listening	75	6.17	0.431
Writing	73	6.08	0.430
Speaking	95	8.00	0.421

Teachers showed stronger assessment knowledge in reading and speaking. Misunderstandings in reading included item independence and test reliability. Listening assessment showed misconceptions in note-taking and phonemic discrimination. In writing, most recognized valid strategies like opinion-based prompts, but misunderstood rubric adaptability and holistic scoring. In speaking, while communicative tasks were favored, misconceptions were noted on task variety and peer interaction protocols.

Correlation analysis showed no significant relationships between demographic variables and LAK levels.

Table 2: Correlation Between Demographic Variables and Overall LAK Scores Source: Author's statistical computation based on survey data

e-ISBN: 978-629-98082-2-0

Variable	Point-Biserial Correlation	p-value
1–5 years of teaching experience	-0.08	0.789 (p > 0.05)
6–10 years of teaching experience	0.08	0.789 (p > 0.05)
Bachelor's degree (BA)	0.40	0.188 (p > 0.05)
Master's degree (MA)	-0.40	0.188 (p > 0.05)
ELT/TEFL/TESOL/CELTA Qualification	-0.22	0.477 (p > 0.05)
General Education Degree	0.22	0.477 (p > 0.05)

No significant correlation was found between the frequency of assessing skills and LAK scores.

Table 3: Correlation Between Skill-Specific LAK and Assessment Frequency Source: Author's statistical computation based on survey data

Variable	Point-Biserial Correlation	p-value
Speaking vs. Speaking LAK	-0.436	0.15 (p > 0.05)
Reading vs. Reading LAK	0.0063	0.98 (p > 0.05)
Writing vs. Writing LAK	-0.440	0.15 (p > 0.05)
Listening vs. Listening LAK	-0.128	0.69 (p > 0.05)

These findings suggest that assessment knowledge in assessing one skill does not transfer to another, and that teaching experience or frequency of assessment does not significantly impact LAK. Prior research (Alderson, 2000; Fulcher & Davidson, 2007; Weir, 2005; Ölmezer-Öztürk, 2018) supports the need for skill-specific professional development to enhance assessment literacy.

4. Conclusion

In conclusion, this study found that English language teachers have a moderate level of assessment knowledge, but their knowledge varies across assessing different language skills. They feel more comfortable assessing reading and speaking, but struggle more with writing and listening. The results in the lack of significant correlations among skill-based LAK levels suggest that assessment knowledge does not automatically transfer from one skill to another. Furthermore, demographic factors such as teaching experience and academic qualifications were not found to significantly influence teachers' assessment knowledge, which highlights the need for structured assessment training rather than relying solely on experience or academic credentials. These findings add to existing professional development programs to help teachers improve their ability to assess language skills effectively.

Based on the findings of this study, it is recommended that professional development programs specifically address teachers' assessment knowledge gaps, particularly in writing and listening assessments. To be effective, these programs should go beyond one-time workshops and incorporate sustained, skill-specific training modules such as in-service certification courses, mentoring programs, or collaborative lesson design. Teacher training programs should also

e-ISBN: 978-629-98082-2-0

integrate comprehensive assessment training within their core curriculum to ensure teachers acquire practical and theoretical competencies across all four language skills. Additionally, schools and institutions should offer ongoing opportunities for professional learning and establish mechanisms to monitor and evaluate the impact of these programs on teachers' Language Assessment Knowledge (LAK) levels and classroom assessment practices.

Acknowledgement

I would like to thank the MEXT Japanese Government Scholarship for supporting my studies as a MEXT scholar. I'm especially grateful to Professor Dr. Yoshinori Watanabe for his kind guidance and to all the teachers who generously participated in this research.

References

Alderson, J. C. (2000). *Assessing reading.* Cambridge University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Brown, H. D. (2003). *Language assessment: Principles and classroom practices*. Pearson Education.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Clapham, C., & Corson, D. (1997). *Language testing and assessment* (Vol. 7). Springer.

Fulcher, G. (2003). *Testing second language speaking*. Pearson Longman.

Fulcher, G. (2012). *Practical language testing*. Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Mertler, C. A. (2003). Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference? *Paper presented at the Annual Meeting of the Mid-Western Educational Research Association*, Columbus, OH.

Mertler, C. A., & Campbell, C. (2005). *Measuring success: Assessment and accountability in education*. Pearson.

Ölmezer-Öztürk, E. (2018). Language assessment knowledge of pre-service and in-service teachers: A review of the literature. *Journal of Language and Linguistic Studies, 14*(1), 1–11. https://www.jlls.org/index.php/jlls/article/view/870

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21–27.

Popham, W. J. (2009). *Assessment literacy for teachers: Faddish or fundamental?* Theory Into Practice, 48(1), 4–11. https://doi.org/10.1080/00405840802577536

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal, 100*(Supplement 2016), 190–208. https://doi.org/10.1111/modl.12308

e-ISBN: 978-629-98082-2-0

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309-327. https://doi.org/10.1177/0265532213480128

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Xu, Y., & Brown, G. T. L. (2017). A systematic review of studies on teachers' assessment literacy: Trends and gaps. *Studies in Educational Evaluation, 55*, 123- 134. https://doi.org/10.1016/j.stueduc.2017.10.006

e-ISBN: 978-629-98082-2-0

P038

Enhancing Assessment Literacy Among Educators: A Mixed-Methods Study on Policy Implementation and Professional Development Practices

Wu Xiaofan

School of Distance Education
Universiti Sains Malaysia

Abstract

This mixed-methods research investigates the policies and practices surrounding assessment literacy for educators, with a specific focus on how professional development programs and institutional policies shape educators' understanding and application of assessment principles, particularly in the context of English language teaching in China. The study employs a sequential explanatory design, beginning with a quantitative phase that surveys 265 educators across diverse educational contexts in Eastern China to measure their levels of assessment literacy, perceived competence, and engagement with assessment-related professional development. After data screening, 256 valid responses were analyzed, revealing key insights into the current state of assessment literacy among English educators. This is followed by a qualitative phase involving in-depth interviews with 12 educators, policymakers, and professional development facilitators to explore the challenges, successes, and gaps in current policies and training programs. The findings highlight three critical issues: (1) While policy implementation has been vigorously promoted, there is a lack of effective mechanisms to assess the learning outcomes of educators in professional development programs. (2) English educators' assessment literacy is primarily confined to traditional classroom practices, with limited familiarity and competence in emerging areas such as machine-based assessment technologies. (3) Professional development courses for assessment literacy are overly uniform, failing to address the diverse needs of educators at different proficiency levels. Additionally, the study examines the role of contextual factors, such as school culture and resource availability, in shaping assessment literacy outcomes. By integrating quantitative and qualitative insights, this research offers evidence-based recommendations for policymakers and educational leaders to design more effective assessment literacy initiatives.

Keywords: assessment literacy, professional development, policy implementation, English language teaching, mixed-methods research

e-ISBN: 978-629-98082-2-0

Introduction

Assessment literacy has emerged as a fundamental competency for educators in the 21st century, encompassing the knowledge, skills, and attitudes necessary to design, implement, and interpret educational assessments effectively (Stiggins, 2002). Contemporary frameworks recognize assessment literacy as extending beyond basic technical knowledge to include professional wisdom, values, conceptions, and ethical responsibilities (Meijer et al., 2023). In the context of English language teaching, language assessment literacy (LAL) becomes particularly crucial as educators must navigate complex linguistic and cultural factors while ensuring accurate measurement of student learning outcomes (Jeong, 2013). Despite growing recognition of its importance, research consistently indicates that many educators lack adequate assessment literacy, leading to ineffective assessment practices and potentially compromised student learning experiences (Xu & Brown, 2016; Jin, 2010).

The implementation of assessment literacy policies and professional development programs has become a priority for educational systems worldwide. However, the effectiveness of these initiatives remains questionable, with limited empirical evidence demonstrating their impact on educator competence and student outcomes (Deluca & Bellara, 2013). Recent research suggests that assessment literacy development requires systematic approaches that incorporate both knowledge and skills components alongside reflective practice and contextual considerations (Pastore & Andrade, 2019). This gap is particularly pronounced in contexts where rapid educational reforms and technological advancements create additional challenges for educators attempting to maintain current assessment practices.

China's educational landscape presents a unique context for examining assessment literacy development. The country's emphasis on high-stakes testing, combined with recent reforms promoting formative assessment and student-centered learning, creates tension between traditional and progressive assessment approaches (Jin, 2010). English language educators in China face additional challenges, including limited exposure to authentic assessment practices and restricted access to professional development opportunities that address their specific needs (Wang et al., 2024). The implementation of China's Standards of English Language Ability (CSE) has further highlighted the need for enhanced assessment literacy among English teachers, as they must align local practices with national standards while maintaining pedagogical effectiveness (Sang, 2023).

The current study addresses these gaps by investigating how policy implementation and professional development practices influence assessment literacy among English educators in Eastern China. Through a mixed-methods approach, this research examines the current state of

e-ISBN: 978-629-98082-2-0

assessment literacy, identifies barriers to effective implementation, and proposes evidence-based recommendations for improvement.

Methods

Research Design

This study employed a sequential explanatory mixed-methods design (Creswell & Plano Clark, 2017), beginning with a quantitative phase to establish baseline understanding of assessment literacy levels, followed by a qualitative phase to explore underlying factors and mechanisms. This approach allowed for comprehensive examination of both the breadth and depth of assessment literacy issues among English educators.

Participants

The quantitative phase involved 265 English educators from diverse educational contexts across Eastern China, including primary schools, secondary schools, and higher education institutions. After data screening for completeness and reliability, 256 valid responses were retained for analysis. Participants ranged in teaching experience from 2 to 30 years, with varying levels of educational background and professional development exposure.

The qualitative phase included 12 purposively selected participants comprising educators (n=8), policymakers (n=2), and professional development facilitators (n=2). Selection criteria included diverse teaching contexts, varying levels of assessment literacy, and willingness to participate in in-depth interviews.

Data Collection

Quantitative data were collected using a validated Assessment Literacy Inventory adapted for the Chinese context (Plake et al., 1993; Zhang & Burry-Stock, 2003). The instrument measured perceived competence across five dimensions, following by the table 1.

Table 1 Dimensions of Assessment Literacy

Dimension	Definition	Sample Competencies
Traditional Assessment Practice	Educators' knowledge and application of conventional assessment methods.	Test administration, marking schemes, interpreting scores, assigning grades.
Assessment Design	The ability to design valid, reliable, and diverse forms of assessments aligned with learning outcomes.	Blueprinting assessments, writing effective items, balancing formative/summative types.
Technology-Enhanced Assessment	Competence in using digital tools and platforms for assessment purposes.	Online quizzes, automated feedback tools, learning analytics, digital rubrics.

e-ISBN: 978-629-98082-2-0

Data Use and Interpretation	The ability to analyze assessment data to inform instructional decisions.	Analyzing student progress, identifying learning gaps, adjusting teaching strategies.
Assessment Ethics and Fairness	Understanding of ethical principles in assessment and ensuring fairness for diverse learners.	Avoiding bias, transparency in grading, providing meaningful feedback.

Qualitative data were gathered through semi-structured interviews lasting 45-60 minutes, conducted in participants' preferred language (Mandarin or English). Interview protocols explored experiences with professional development, policy implementation challenges, and suggestions for improvement.

Data Analysis

Quantitative data were analyzed using descriptive statistics, correlation analysis, and multiple regression to identify patterns and predictors of assessment literacy. Qualitative data underwent thematic analysis following Braun and Clarke's (2006) framework, with coding conducted independently by two researchers to ensure reliability.

Results and Discussion

Current State of Assessment Literacy

The quantitative findings revealed moderate levels of assessment literacy among English educators (M = 3.42, SD = 0.67 on a 5-point scale). Participants demonstrated highest competence in traditional assessment practices, including test administration and basic result interpretation, both average score above 3.5, which means the good level of practice and understanding. However, significant gaps emerged in advanced areas such as assessment design and technology-enhanced assessment, as the figure 1 shows, below 3 points.

e-ISBN: 978-629-98082-2-0

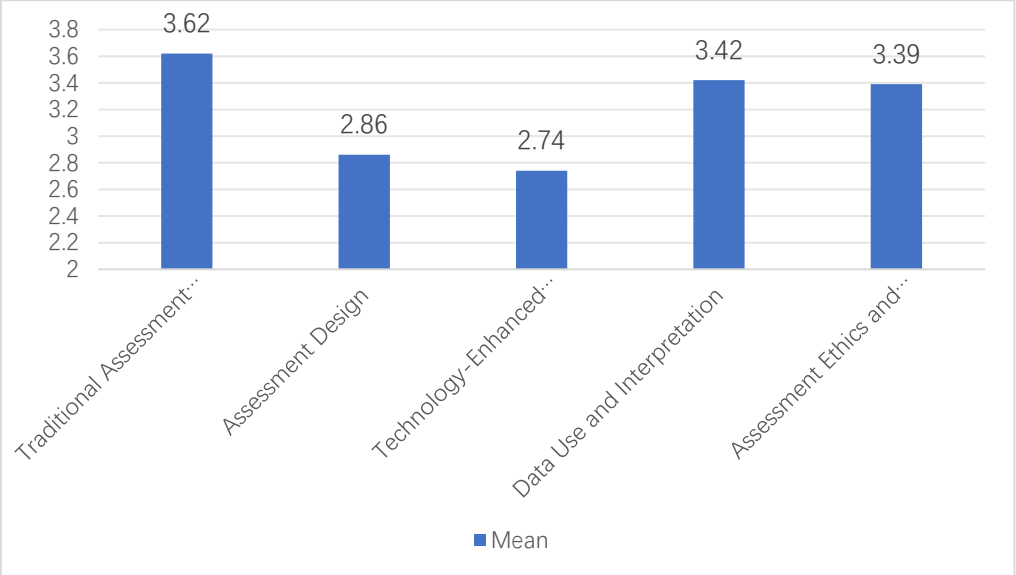


Figure 1 Average score of each dimension

These results align with international research indicating that educator assessment literacy remains underdeveloped globally (Pastore & Andrade, 2019; Volante & Fazio, 2007). The particularly low scores in technology-enhanced assessment reflect the challenges educators face in adapting to digital assessment tools and machine-based evaluation systems, a concern that has become increasingly relevant in post-pandemic educational contexts (O'Leary et al., 2024). Similar patterns have been observed in other Chinese contexts, where English teachers demonstrate stronger competence in traditional assessment methods but struggle with innovative assessment approaches (Liu & Jin, 2024).

Policy Implementation Challenges

The qualitative analysis revealed three primary challenges in policy implementation. First, there is a significant disconnect between policy intentions and practical implementation. While assessment literacy policies have been widely promoted, educators report limited understanding of how to translate policy requirements into effective classroom practices. Participant 5 noted, *"We receive many documents about assessment reform, but little guidance on actual implementation."*

Second, professional development programs lack effective evaluation mechanisms. Educators participate in training sessions but receive minimal feedback on learning outcomes or practical application. This finding supports the quantitative results showing weak correlations between professional development participation and assessment literacy scores ($r = 0.23$, $p < 0.05$).

e-ISBN: 978-629-98082-2-0

Third, institutional support varies significantly across contexts. Schools with stronger administrative support and resource availability demonstrate higher levels of assessment literacy implementation, suggesting that contextual factors play crucial roles in policy success.

Professional Development Effectiveness

The study identified substantial limitations in current professional development approaches. Most programs follow a "one-size-fits-all" model that fails to address diverse educator needs and experience levels (Popham, 2018). Novice educators require foundational knowledge, while experienced educators need advanced training in emerging assessment technologies and innovative practices. Research suggests that effective assessment literacy development requires differentiated approaches that consider teachers' prior knowledge, contextual factors, and specific learning objectives (Deluca et al., 2016).

Additionally, professional development programs rarely include practical application components or follow-up support. Educators report attending workshops but struggling to implement new concepts without ongoing guidance and feedback (Yan et al., 2018). This gap between training and application may explain the modest correlation between professional development participation and assessment literacy competence. Studies in similar contexts have shown that sustained professional learning communities and mentorship programs are more effective than isolated training sessions (Jamil & Hamre, 2018).

Recommendations for Improvement

Based on the integrated findings, several recommendations emerge for enhancing assessment literacy initiatives:

Differentiated Professional Development: Design tiered training programs that address varying levels of educator expertise and experience. Novice educators should receive comprehensive foundational training, while experienced educators require specialized modules focusing on advanced assessment techniques and emerging technologies.

Technology Integration: Incorporate systematic training on machine-based assessment tools and digital evaluation platforms. This should include hands-on practice with software applications and guidance on interpreting automated assessment results.

Evaluation Mechanisms: Establish robust systems for evaluating professional development effectiveness, including pre- and post-training assessments, classroom observation protocols, and long-term follow-up surveys to measure sustained implementation.

e-ISBN: 978-629-98082-2-0

Contextual Support: Develop school-based support systems that include administrative backing, resource allocation, and peer collaboration opportunities. This may involve training school leaders to understand and support assessment literacy initiatives.

Continuous Learning Communities: Create ongoing professional learning communities where educators can share experiences, discuss challenges, and collaborate on assessment innovation. These communities should be supported by expert facilitators and connected to broader professional networks.

Conclusion

This mixed-methods study provides comprehensive insights into the current state of assessment literacy among English educators in China and identifies critical areas for improvement in policy implementation and professional development practices. The findings demonstrate that while assessment literacy policies have been widely promoted, significant gaps remain in effective implementation and educator competence development.

The study's three key findings – inadequate evaluation mechanisms, limited technology competence, and uniform professional development approaches – highlight systemic issues that require coordinated responses from policymakers, educational leaders, and professional development providers. The recommendations offered provide evidence-based directions for creating more effective assessment literacy initiatives that address diverse educator needs and promote sustained implementation.

Future research should examine the long-term impact of differentiated professional development programs and investigate the role of cultural factors in shaping assessment literacy development. Additionally, comparative studies across different educational contexts could provide valuable insights into universal versus context-specific factors influencing assessment literacy outcomes.

The implications of this research extend beyond the Chinese context, offering insights relevant to international efforts to enhance educator assessment literacy. As educational systems worldwide grapple with similar challenges related to assessment reform and professional development effectiveness, the findings and recommendations presented here contribute to the global conversation on improving educational assessment practices.

e-ISBN: 978-629-98082-2-0

Acknowledgments

The authors gratefully acknowledge the educators, policymakers, and professional development facilitators who participated in this study. This research has no funding.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). Sage Publications.

Deluca, C., & Bellara, A. (2013). The current state of assessment education: Aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education*, 64(4), 356-372. https://doi.org/10.1177/0022487113488144

Deluca, C., Lapointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28(3), 251-272. https://doi.org/10.1007/s11092-015-9233-6

Jamil, F. M., & Hamre, B. K. (2018). Teacher reflection in the context of an online professional development course: Applying principles of cognitive science to promote teacher learning. *Action in Teacher Education*, 40(2), 220-236. https://doi.org/10.1080/01626620.2018.1424051

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers? *Language Testing*, 30(3), 345-362. https://doi.org/10.1177/0265532213480334

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing*, 27(4), 555-584. https://doi.org/10.1177/0265532210384550

Liu, Y., & Jin, Y. (2024). Understanding university English instructors' assessment literacy: A formative assessment perspective. *Language Testing in Asia*, 14(1), 1-22. https://doi.org/10.1186/s40468-024-00323-y

Meijer, K., Kuijper, H., Bakker, A., & Admiraal, W. (2023). Teachers' conceptions of assessment literacy. *Teachers and Teaching*, 29(3), 289-309. https://doi.org/10.1080/13540602.2023.2190091

O'Leary, M., Cui, V., & French, S. (2024). The key competencies and components of teacher assessment literacy in digital environments: A scoping review. *Teaching and Teacher Education*, 138, 104-119. https://doi.org/10.1016/j.tate.2024.104389

e-ISBN: 978-629-98082-2-0

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, 84, 128-138. https://doi.org/10.1016/j.tate.2019.05.003

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, 12(4), 10-12. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Popham, W. J. (2018). *Assessment literacy for educators in a hurry*. ASCD.

Sang, Y. (2023). China's Standards of English Language Ability: Voice from English teachers at Chinese universities. *SAGE Open*, 13(4), 1-15. https://doi.org/10.1177/21582440231205434

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10), 758-765. https://doi.org/10.1177/003172170208301010

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education*, 30(3), 749-770. https://doi.org/10.2307/20466661

Wang, L., Zhang, H., & Chen, Y. (2024). Exploring Chinese university English teachers' language assessment literacy: A mixed-method study. *Asia Pacific Journal of Education*, 44(3), 456-472. https://doi.org/10.1080/02188791.2024.2354684

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162. https://doi.org/10.1016/j.tate.2016.05.010

Yan, Z., Li, Z., Pastore, S., Gotch, C. M., & Yang, M. (2018). An investigation into assessment literacy of primary school teachers in China. *Studies in Educational Evaluation*, 58, 27-38. https://doi.org/10.1016/j.stueduc.2018.05.004

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, 16(4), 323-342. https://doi.org/10.1207/S15324818AME1604_4

e-ISBN: 978-629-98082-2-0

P041

Gamification in Formative Assessment for Non-Majors: Engagement and Challenges

Thuy Duy T. Pham[1], **Nha Phuong T. Nguyen**[2]

[1] Faculty of Foreign Languages
Tra Vinh University, Vietnam
E-mail: [1]*thuyduy@tvu.edu.vn*

[2]Faculty of Foreign Languages
Tra Vinh University, Vietnam
E-mail: [2]*nguyenthinhaphuong@tvu.edu.vn*
**corresponding author:* [1]*thuyduy@tvu.edu.vn*

Abstract

With the rapid development of educational technology, gamification has attracted widespread application in schools as a potent tool for engaging students and evaluating teachers' instructional quality. This implies that through game-based online platforms teachers are forced to reconsider their teaching techniques as they stimulate a dynamic and participative classroom approach. This study aims to explore the effects of using formative tests via a gamification platform like Kahoot! on student engagement and challenges. The research was conducted with a group of 86 non-English major students who were taking general English courses. The students were guided and then completed five in-class quizzes as part of their formative assessment. To collect data, a questionnaire was adapted from the framework that assesses engagement from four perspectives: affective, behavioral, cognitive, and agentic engagement. Besides, the semi-structured interview was conducted in order to find out some challenges. Findings show that Kahoot! strongly promoted affective and behavioral engagement, as students felt motivated and actively involved during quizzes. However, cognitive engagement was sometimes limited by time constraints, technical issues, and distracting visuals. The study suggests that while gamified tools like Kahoot! boost emotional involvement and participation, they should be used alongside other assessment methods such as discussions or writing tasks. Institutions and educators are encouraged to provide students with technical assistance, and adjust quiz design to reduce distractions, maximize learning benefits and support long-term student engagement.

Keywords: formative assessment; gamification; Kahoot!; non-majors; engagement

e-ISBN: 978-629-98082-2-0

1. Introduction

Gamification has emerged as a prominent technique in language education. Gamification not only enhances excitement and engagement among learners but also serves as an excellent instrument for educators to assess and refine their teaching practices in alignment with a learner-centred approach (Göksün & Gürsoy, 2019).

In EFL classrooms, maintaining active engagement and motivation of students, especially non-English majors, is a considerable challenge. Many prior studies have shown that traditional methods are not effective enough in motivating learners and promoting interaction in the classroom (Bouwmeester et al., 2019). Gamification not only brings about a sense of achievement and excitement (Li et al., 2022) but also supports learners in developing thinking skills, multitasking ability, and increasing self-confidence through game elements (Ding et al., 2018). Although gamified platforms have been explored in various educational contexts (Zainuddin et al., 2020), their application as formative assessment tools in English classrooms for non-English majors at the university level, particularly in Vietnam, remains under-researched. Moreover, though Göksün and Gürsoy (2019) reported such difficulties based on students' qualitative feedback, the challenges of students using kahoot! as a formative assessment tool have not been sufficiently addressed. Zainuddin et al. (2020) also recommended that future research should focus on these practical challenges in gamified formative assessment.

This study explores non-English majors' perceptions of using Kahoot! for formative assessment in General English classes, focusing on its impact on student engagement and the challenges they experienced during its use, guided by two research questions: (1) How do students perceive the impact of gamified formative assessment on their engagement? (2) What challenges do students face when participating in gamified formative assessment activities?

2. Methods

2.1. Participants

The population was a cohort of 86 non-English majors at intermediate level from three General English classes at a Vietnamese University from differerent majors. The students were selected using a convenience sampling method based on their enrollment and availability during the study period.

2.2. Research design

e-ISBN: 978-629-98082-2-0

This study employed a mixed-method design, combining quantitative and qualitative data to investigate non-English majors' engagement with gamified formative assessments using Kahoot! During 15 weeks, the students took five formative assessment tests on Kahoot!. Students were invited to complete the questionnaire. Additionally, semi-structured interviews were performed with 12 chosen students to obtain more profound insights into their perceptions.

2.3. Data collection and analysis

The questionnaire comprised 14 items utilising a five-point Likert scale, corresponding to four components of the engagement framework: emotive, behavioural, cognitive, and agentic adapted from Reeve and Tseng (2011). The questionnaire was validated by two experts in applied linguistics and piloted with a small group of students (n = 10) for content validity and underwent minor revisions based on the results of the pilot.

For collecting the data, the participants were invited to complete the questionnaire anonymously via Google Forms. For interviews, twelve students were selected and invited to answer the questions related to the challenges. Quantitative data were analyzed using SPSS 20 while interview data were thematically analyzed.

3. Results and Discussion

3.1 The results from the questionnaire

Table 1. The reliability analysis

Dimension	**Items**	**Cronbach's Alpha**
Affective Engagement	A1, A2	0.797
Behavioral Engagement	B1, B2, B3	0.872
Cognitive Engagement	C1, C2, C3, C4, C5	0.951
Agentic Engagement	AG1, AG2, AG3, AG4	0.952

The reliability analysis results demonstrated that all engagement subscales showed suitable to excellent internal consistency with alpha values between 0.797 (affective) and 0.952 (agentic). These results confirm the instrument effectively measured the targeted engagement constructs in EFL learner populations.

e-ISBN: 978-629-98082-2-0

Table 2. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**.928**
Bartlett's Test of Sphericity	Approx. Chi-Square	1211.579
	df	78
	Sig.	**.000**

Table 3. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	9.213	70.867	70.867	9.213	70.867
2	.936	7.198	78.064	.936	7.198
3	.616	4.742	82.806	.616	4.742
4	.515	3.960	86.767	.515	3.960
5	.347	2.673	89.439		
6	.332	2.553	91.993		
7	.302	2.322	94.315		
8	.185	1.426	95.741		
9	.157	1.211	96.952		
10	.119	.912	97.864		
11	.107	.825	98.689		
12	.088	.675	99.365		
13	.083	.635	100.000		

e-ISBN: 978-629-98082-2-0

Table 4. Pattern Matrix[a]

	Component			
	1	2	3	4
AG2	.937			
AG4	.881			
AG3	.875			
AG1	.629			
C2		.782		
C1		.773		
C3		.718		
C5		.593		
B3			.896	
B2			.752	
B1			.617	
A2				.927
A1				.754

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.[a]
a. Rotation converged in 6 iterations.

To examine the construct validity of the engagement scale, exploratory factor analysis was conducted using principal component extraction and Promax rotation. The Kaiser-Meyer-Olkin (KMO) measure was 0.928, indicating excellent sampling adequacy, and Bartlett's Test of Sphericity was significant ($\chi^2(78) = 1211.579$, $p < .001$), confirming that the data were appropriate for factor analysis. Four factors were extracted based on eigenvalues greater than 1, accounting for 86.77% of the total variance. The rotated pattern matrix revealed clear factor loadings: agentic engagement items (AG1–AG4) loaded on Factor 1, cognitive engagement items (C1–C5) on Factor 2, behavioral engagement items (B1–B3) on Factor 3, and affective engagement items (A1–A2) on Factor 4. These results support the construct validity of the scale and align with the theoretical model of student engagement.

Table 5. Descriptive Statistics for Engagement Types

Engagement Type	**Mean**	**Std. Deviation**	**N**
Affective	4.12	1.09	86
Behavioral	4.11	0.95	86
Cognitive	3.96	1.09	86
Agentic	3.82	1.05	86

e-ISBN: 978-629-98082-2-0

The means of affective, behavioral, cognitive and agentic engagement underwent a repeated-measures ANOVA to assess if their mean scores showed statistically significant differences. Table 5 shows that among the four dimensions affective engagement (M = 4.12, SD = 1.09) and behavioral engagement (M = 4.11, SD = 0.95) received the highest positive impact from gamification during formative assessment. The ratings for cognitive (M = 3.96, SD = 1.09) and agentic engagement (M = 3.82, SD = 1.05) were lower than the other dimensions.

Table 6. Mauchly's Test of Sphericity[a]

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]
					Greenhouse-Geisser
Engagement_Sub	.848	13.967	5	.016	.903

Table 7. Repeated-Measures ANOVA Summary

Source	df	F	*p*	Partial η²
Engagement_Sub	2.709	6.49	**0.001**	0.071
Error	230.256	—	—	—

Mauchly's test showed that the sphericity assumption was violated with $\chi^2(5) = 13.967$ and $p = 0.016$; therefore Greenhouse-Geisser correction was applied ($\varepsilon = 0.903$). The repeated-measures ANOVA demonstrated a statistically significant main effect of engagement type, $F(2.709, 230.256) = 6.49$, $p = 0.001$, partial $\eta^2 = 0.071$ (Table 7) which showed that gamification effects on perception varied between four engagement domains. Pairwise comparisons exist to determine which particular engagement dimensions have significant differences after the repeated-measures ANOVA detected an overall significant effect.

Table 8. Pairwise Comparisons of Engagement Types (Bonferroni-Adjusted)

Comparison (I–J)	Mean Difference	Std. Error	*p*-value	95% CI (Lower, Upper)	Significant
Affective vs. Behavioral	0.004	0.084	1.000	[–0.224, 0.232]	No
Affective vs. Cognitive	0.157	0.082	0.357	[–0.065, 0.379]	No
Affective vs. Agentic	**0.298**	0.093	**0.011**	[0.048, 0.549]	**Yes**
Behavioral vs. Cognitive	0.153	0.064	0.113	[–0.020, 0.326]	No
Behavioral vs. Agentic	**0.295**	0.077	**0.002**	[0.086, 0.503]	**Yes**
Cognitive vs. Agentic	0.141	0.071	0.306	[–0.052, 0.335]	No

The Bonferroni-adjusted pairwise comparisons in Table 8 reveal that participants found gamification in formative assessment to have a stronger effect on affective engagement compared to agentic engagement (M difference = 0.30, p = 0.011, 95% CI [0.048, 0.549]). Behavioral engagement received significantly higher ratings than agentic engagement

e-ISBN: 978-629-98082-2-0

according to the pairwise comparisons (M difference = 0.30, p = 0.002, 95% CI [0.086, 0.503]). The data demonstrates that affective and behavioral engagement types do not show any statistically significant differences along with the absence of any differences between cognitive and other engagement types (all p > .05).

The findings indicate that Kahoot! and Quizizz as gamified formative assessment tools produce significant improvements in students' affective and behavioral engagement while showing limited effects on cognitive and agentic engagement. The repeated-measures ANOVA analysis revealed statistically significant differences between the engagement dimensions because affective (M = 4.12) and behavioral (M = 4.11) scores surpassed cognitive (M = 3.96) and agentic (M = 3.82) scores. The findings of this study confirm the results found in previous research. For instance, the implementation of Kahoot! in chemistry classes led to improved student engagement according to Al Ghawail and Yahia (2022) research while Hoang (2024) observed strong emotional responses from students who used Quizizz. The current study along with previous investigations demonstrates that cognitive and agentic improvements remain limited. According to Munawir and Hasbi (2021), gamification methods increase motivation and performance levels yet fail to develop deep thinking abilities and learner independence.

3.2 The results from the interviews

While Kahoot! brings about potential benefits, several key challenges were revealed. Firstly, some students found Kahoot!'s platform characterized by bright colors, sound effects, and fast pace distracting, which occasionally took attention away from the content, which is in alignment with Annetta (2010). Secondly, the limited time for answering questions made students select answers quickly before they fully understood the questions as they wanted to avoid missing their opportunity to respond. This corresponds with the findings of Bicen and Kocakoyun (2018), showing that time-based tasks could affect thoughtful engagement. Another problem related to technical issues like low internet connection or phone battery disrupted participation. This issue is often overlooked in gamification research but is crucial for effective implementation (Denisova & Cairns, 2015). Moreover, the students provided some meaningful suggestions for improving gamified assessment. These included "extending time for difficult questions", "adding explanations after responses", "enabling quiz reviews", and "combining Kahoot! with other assessment formats" (e.g., discussion or writing). In general, it is important for the teachers to pay attention to aligning its design with pedagogical goals, ensuring accessibility, and supporting meaningful understanding.

4. Conclusion

Students experienced the greatest positive impact on their affective and behavioral engagement when gamification was implemented in formative assessment. The results

e-ISBN: 978-629-98082-2-0

indicate that Kahoot! and other gamified tools boost student emotional engagement and participation. However, some technical problems and cognitive overload resulting from flashy interface features as obstacles, which highlight the necessity of creating learning support systems instead of distracting elements. Educational institutions should establish complete technical preparedness alongside measures to reduce elements, which create distractions. Educators should combine gamified tools with other assessing methods such as discussions and written activities to maintain engagement while achieving meaningful learning outcomes. The educational value of these tools could improve through design modifications, which extend time periods and provide feedback and adjust competitive elements and consider different kinds of awards.

Acknowledgement

We would like to express our sincere thanks to all the participants who spend their time to completing the questionnaire and expressing their opinion for the interviews. Our heartfelt gratitude also goes to Tra Vinh University (TVU) for the valuable support and resources that made this research possible.

References

Al Ghawail, E. A., & Yahia, S. B. (2022). Using the e-learning gamification tool Kahoot! to learn chemistry principles in the classroom. *Procedia Computer Science, 207*, 2667–2676.

Annetta, L. A. (2010). The "I's" have it: A framework for serious educational game design. *Review of General Psychology, 14*(2), 105–113.

Bicen, H., & Kocakoyun, S. (2018). Perceptions of students for gamification approach: Kahoot as a case study. *International Journal of Emerging Technologies in Learning, 13*(2), 72–93.

Bouwmeester, R. A., de Kleijn, R. A., van den Berg, I. E., ten Cate, O. T. J., van Rijen, H. V., & Westerveld, H. E. (2019). Flipping the medical classroom: Effect on workload, interactivity, motivation and retention of knowledge. *Computers & Education, 139*, 118–128.

Denisova, A., & Cairns, P. (2015). Adaptation in digital games: The effect of challenge adjustment on player performance and experience. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (pp. 97–101). ACM.

Ding, L., Er, E., & Orey, M. (2018). An exploratory study of student engagement in gamified online discussions. *Computers & Education, 120*, 213–226.

Göksün, D. O., & Gürsoy, G. (2019). Comparing success and engagement in gamified learning experiences via Kahoot and Quizizz. *Computers & Education, 135*, 15–29.

e-ISBN: 978-629-98082-2-0

Hoang, T. L. G. (2024). EFL learner engagement in gamified formative assessment: A perception study on Quizziz. *Hue University Journal of Science: Social Sciences and Humanities, 133*(6B), 23–44.

Li, X., Xia, Q., Chu, S. K. W., & Yang, Y. (2022). Using gamification to facilitate students' self-regulation in e-learning: A case study on students' L2 English learning. *Sustainability, 14*(12), 7008.

Munawir, A., & Hasbi, N. P. (2021). The effect of using Quizizz to EFL students' engagement and learning outcome. *English Review: Journal of English Education, 10*(1), 297–308.

Reeve, J., & Tseng, C. M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology, 36*(4), 257–267.

Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system. *Computers & Education, 145*, 103729.

e-ISBN: 978-629-98082-2-0

P043

A Study on the Reliability of Scoring English Essays by Advanced English Learners Using ChatGPT

Jungyeon Koo

Department of English Language and Literature, Seoul National University, Korea.
(E-mail: [1]christy9r@snu.ac.kr)
**corresponding author: [1]christy9r@snu.ac.kr*

Abstract

In an era defined by the Fourth Industrial Revolution, technological innovations—particularly those driven by artificial intelligence (AI) and big data—are transforming diverse sectors. One prominent trend in English language testing and assessment is the integration of AI, especially in automating evaluation processes. Large Language Models (LLMs), in particular, are noted for their consistent scoring performance. This study aims to compare the scoring outputs of two ChatGPT versions—3.5 and 4.0 (Omni mini)—and evaluate their alignment with human raters. A total of 150 university students were instructed to write English essays online within 30 minutes, responding to two different types of prompts. Four human raters evaluated the essays using both analytic and holistic scoring rubrics. A strong correlation was observed between the holistic and analytic scoring results. Essays that exhibited large discrepancies between human scoring and automated essay scoring (AES) were found to differ significantly in content features. This research highlights the importance of narrowing the gap between human scoring and AES, which may assist instructors in recognizing ChatGPT as a dependable tool for educational assessment. Additionally, the findings underscore the potential of ChatGPT as an effective tool for evaluating English writing performance. This study is meaningful in that it provides practical insights and guidelines for improving the efficiency and reliability of automated essay scoring systems in academic settings.

Keywords: automated essay Scoring, ChatGPT, large language models, advanced English learners of English, Reliability

1. Introduction

In the era of the Fourth Industrial Revolution, technological innovations and changes driven by artificial intelligence (AI) and big data are transforming various fields. Specifically, in English language assessment, large language models (LLMs) are noted for their consistency in scoring, especially with the advancement of models like GPT. Thus, this study aims to explore the

e-ISBN: 978-629-98082-2-0

application of AI tools based on LLM for scoring English essays in order to determine the high accuracy of score calibration.

The growing recognition of the importance of writing, coupled with the high cost and time demands of reliable and valid human grading, has increased the demand for more rapid assessment procedures. Consequently, this has accelerated the development of Automated Essay Scoring (AES) systems. Traditionally, these systems have relied on a combination of computational linguistics, statistical modeling, and natural language processing (NLP) (Shermis & Burstein, 2013), prior to the emergence of ChatGPT in 2022. Significant progress has been made by integrating cutting-edge technologies such as NLP, machine translation, deep learning (DL), and speech recognition and synthesis (Baidoo-Anu & Ansah, 2023; Kaplan, 2015; Schwab, 2018). More recently, with the advent of Generative AI (GAI) tools such as ChatGPT and Gemini, profound changes have occurred across various sectors, including information and communication, healthcare, logistics, and education (Baidoo-Anu & Ansah, 2023; Gupta et al., 2023; Yu et al., 2024).

A recent noteworthy trend in English language testing and assessment is the application of AI, particularly in the automation of language assessment. Specifically, large language models (LLMs), such as those developed by OpenAI, have demonstrated reliable scoring capabilities. Therefore, this study compares the scoring outputs of two ChatGPT versions—3.5 and 4.0 (Omni) mini)—and evaluates their agreement with human scorers.

2. Methods

A total of 150 undergraduate and graduate students at several Korean universities were recruited online. The participants came from diverse academic majors and had an average TOEFL iBT score of 98. They were instructed to write English essays within 30 minutes on two prompts selected from the TOEFL11 corpus developed by Educational Testing Service (ETS). Four trained human raters scored the essays using both analytic and holistic rubrics.

For automated scoring, an LLM-based essay evaluation tool was developed using ChatGPT's API. ChatGPT 3.5 and ChatGPT 4-O (Omni mini) were employed for both holistic and analytic scoring. Zero-shot prompting—without prior examples—was used. Three metrics were analyzed: reliability, accuracy, and consistency of ChatGPT's scoring. Data analysis was conducted using SPSS and the Facets Program. Correlations between holistic and analytic scoring were examined, and the agreement between GPT-generated scores and human ratings was evaluated. Reliability was assessed by having ChatGPT rescore the same essays three

e-ISBN: 978-629-98082-2-0

times, and internal consistency was measured using Intraclass Correlation Coefficient (ICC) and one-way repeated measures ANOVA.

To implement zero-shot prompting, a web-based tool called the Essay Evaluation Support System (EESS) was utilized (see Figure 1).

Figure 1: Essay Evaluation Support System (EESS) for zero prompting

The EESS is produced by linking API (Application Program Interface) for each version of GPT using a back-end framework, “nodejs” based on Java script, one of the representative web-based languages. The reason the EESS was developed with the Java script is that this evaluator system is web-based. The EESS consists of five parts: (1) home, (2) essay management, (3) evaluation criteria management, (4) evaluation plan and assessment, and (5) analytic statistical management

Evaluation types were selected by the user, with options for GPT-3.5 or GPT-4o-mini. GPT-3.5, containing approximately 1.75 billion parameters, processes data faster but is generally less precise than GPT-4, which has over 1 trillion parameters. GPT-4o, OpenAI's flagship model, offers improved speed, reduced costs, and multimodal capabilities. GPT-4o-mini provides faster processing while maintaining high performance on targeted tasks.

Figure 2 presents the layout of the EESS interface, which includes uploaded essays and scoring models.

e-ISBN: 978-629-98082-2-0

Figure 2: The Layout of Essay Evaluation Support System (EESS)

1	MACHINE	GPT3	GPT3 테스트 (1회차)	30 / 30	2023-10-14	해당사항 없음
2	HUMAN	구정연	30개 인간평가_구정연	30 / 30	2023-10-14	엑셀업로드
3	HUMAN	지예도	30개 인간평가 지예도	30 / 30	2023-10-15	엑셀업로드
4	HUMAN	조미라	30개 인간평가 조미라	30 / 30	2023-10-15	엑셀업로드
5	HUMAN	김무현	30개 인간평가 공진솔	30 / 30	2023-10-15	엑셀업로드
6	MACHINE	GPT4	GPT4 신뢰도 테스트1차	2 / 30	2023-12-27	해당사항 없음
7	MACHINE	GPT4	gpt4 신뢰도테스트 2차	0 / 0	2023-12-27	해당사항 없음
8	MACHINE	GPT4	gpt4 신뢰도테스트 3차	0 / 0	2023-12-27	해당사항 없음

3. Results and Discussion

Automated scoring with ChatGPT 3.5 and 4-o-mini indicated that analytic scoring yielded higher reliability and accuracy than holistic scoring. Essays with large discrepancies between human and GPT scores typically varied in content-focused components. Specifically, for surface-level criteria such as lexical complexity, mechanical accuracy, and sentence structure, ChatGPT's repeated scores demonstrated high reliability and consistency. However, for content-driven aspects like task fulfillment, paragraph coherence, and inter-sentential cohesion, the models exhibited lower reliability.

These findings suggest that: 1) Analytic rubrics yield more accurate AI-based scoring than holistic rubrics for advanced Korean EFL learners; 2) LLM-based assessment tools are more effective for evaluating surface features than content-related criteria; and 3) GPT 4-o-mini demonstrates stronger alignment with human scoring than GPT 3.5.

Table 1. Descriptive Statistics of ChatGPT- 3.5 and 4-o-mini, and Human

		Mean	***SD***	***Var***	***Max***	***Min***
Human	H	2.75	0.44	0.195	4	2
	A	2.91	0.60	0.36	4	1
GPT 3.5	H	2.75	0.44	0.195	4	2
	A	3.05	0.42	0.18	4	1
	H	2.27	0.48	0.23	3	1

e-ISBN: 978-629-98082-2-0

GPT 4-o-mini	A	2.27	0.51	0.26	4	1

Note: H means holistic, A means analytic scoring

The table 1 shows that GPT 3.5 scores showed the same mean, standard deviation, variation and even maximum and minimum score with human scores in holistic scoring. GPT 4-0-mini demonstrated the lower mean and standard deviation than GPT 3.5. This means that GPT 4-0-mini scores more severely and also shows wider score variation than GPT 3.5 and human do in holistic scoring. Also, GPT 4-0-mini demonstrated the lowest mean score compared to the other scorer.

In contrast, GPT 3.5 scores showed the same mean, standard deviation, variation and even maximum and minimum score with human scores in holistic scoring. In analytic and holistic scoring, GPT 3.5 evaluated essays more leniently than GPT 4-0-mini in terms of mean scores (holistic: GPT 3.5. vs. GPT 4-0-mini: 2.75 vs 2.27, analytic: human vs. GPT 3.5. & GPT 4-0-mini: 3.05 vs. 2.27). This result might be due to GPT4-0-mini can detect errors more severely because GPT 4-0-mini has more parameters and trained with more internet-based contents.

The reliability between holistic rubric and analytic rubric are shown in Table 2.

Table 2. Inter-rater reliability between holistic and analytic scoring in two scoring models

	Human	GPT 3.5	GPT 4-o-mini
ICC	0.76	0.25	0.94
Pearson Correlation	0.76	0.31	0.94

Note: ICC means Intraclass correlation coefficient.

GPT 4-o-mini demonstrated the highest consistency between holistic and analytic scoring (ICC and Pearson r = 0.94), surpassing both human raters and GPT 3.5. GPT 3.5 showed the lowest internal consistency. These results indicate that GPT 4-o-mini is more consistent and reliable across scoring dimensions. Table 3 compares the correlation of each GPT model with human analytic scores.

Table 3. Pearson Correlation with Human Ratings in analytic scoring

Model	Pearson Correlation
GPT-3.5 vs Human	-0.494
GPT-4-o-mini vs Human	0.995

e-ISBN: 978-629-98082-2-0

GPT 4-o-mini exhibited a very high positive correlation with human analytic scores, while GPT 3.5 showed a negative correlation. This suggests GPT 4-o-mini more closely replicates human judgment and that GPT 3.5 may require additional prompting strategies to improve performance.

4. Conclusion

The findings suggest that ChatGPT 4-o-mini demonstrates lower scoring bias and higher consistency compared to human raters. This study contributes meaningful insights into efficient, accurate scoring for advanced EFL learners, providing practical guidance for integrating LLM-based AES systems into educational assessment.

Acknowledgement

I would like to express sincere gratitude to the anonymous reviewers of the Asian Association for Language Assessment (AALA). Special thanks are also extended to Seoul National University for its academic support and research environment that made this study possible.

References

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI, 7(1), 52-62. http://dx.doi.org/10.2139/ssrn.4337484

Gupta, B., Mufti, T., Sohail, S. S., & Madsen, D. Ø. (2023). ChatGPT: A brief Narrative Review. Cogent Business & Management, 10, https://doi.org/10.1080/23311975.2023.2275851 University Press.

Kaplan, J. (2015). Humans Need not Apply: A guide to Wealth and Work in the Age of Artificial Intelligence. Yale University Press.

Schwab, K. (2018). *Shaping the future of the fourth industrial revolution*. Currency.

Shermis, M. D. & Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routeledge.

Yu, J. H., Chauhan, D., Iqbal, R. A., & Yeoh, E. (2024). Mapping Academic Perspectives on AI in Education: Trends, Challenges, and Sentiments in Educational Research (2018-2024). Education Tech Research Dev. https://doi.org/10.1007/s11423-024-10425-2

e-ISBN: 978-629-98082-2-0

P047

A.I. or Human? A Study of AI-Powered Speech in Tertiary Level Listening Test for EFL Learners

***Khairi Fakhri Fazil[1], *Nur Ehsan Mohd Said[2], Pham Ngoc Bao Tram[3] and Zeng Yijing[4]**

[1,3,4]Centre for Shaping Advanced & Professional Education (UKMShape),
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, MALAYSIA.
[2]Faculty of Education,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, MALAYSIA.

(E-mail: [1]khairifakhri@ukm.edu.my, [3]phngbaotram@ukm.edu.my, [4]zengyijing@ukm.edu.my)
**Corresponding author: [2]nurehsan@ukm.edu.my*

Abstract

Artificial intelligence (AI) has become increasingly prominent in enhancing the accessibility of language testing, despite underexplored reporting of AI-powered speech in listening tests. This study aims to investigate the human-likeness of an AI-powered speech for tertiary-level English as a Foreign Language (EFL) students in listening test and explore the extent of it in listening assessment demands. The two-phase study employed the Fuzzy Delphi Method (FDM) and interviews for further expert insights. Evaluation criteria were identified, and 7 expert informants in CEFR-aligned assessments were selected. The first phase of this study explores the consensus agreement between experts and the defuzzification value for the aspects of 1) Word Stress, 2) Rhythm, 3) Intonation, 4) Pronunciation Accuracy, 5) Enunciation and 6) Intelligibility in rivaling human speech for listening tests demands. A threshold (α-cut) of ≤ 0.5 and a minimum of 70% expert agreement were used as benchmarks of consensus. Quantitative analysis ranked Pronunciation Accuracy, Intelligibility, Enunciation, and Word Stress highest in defuzzified value (A), but only Enunciation and Word Stress met consensus. Accuracy and Intelligibility showed wider rating dispersion despite their high approval value as opposed to Rhythm and Intonation. Follow-up interviews confirmed that AI-generated speech is mainly human-comparable with anomalies in Rhythm and Intonation at minimal level due to pacing inconsistencies. Despite this, expert informants considered it suitable for testing. These initial findings are grounded on a single stimulus; whereby future cycles will observe a broader range of variations in AI-powered speech.

Keywords: AI-powered speech; language testing; listening test; EFL; tertiary education

e-ISBN: 978-629-98082-2-0

1. Introduction

The increasing advances in the use of artificial intelligence (AI) in language assessment has garnered progressive interest in the recent decade, in tandem with the need for more accessible and scalable applications in educational landscapes. Despite growing interest in AI applications for language testing, empirical research on the use of AI-generated speech in EFL listening tests remains limited. While AI-driven writing evaluations such as automated scoring systems are now made available for assessment purposes (Lu & Cutumisu 2021; Susanti et al. 2023), studies on the application of AI in listening assessments particularly in evaluating AI-driven stimulus in addressing listening assessment demands in English as a Foreign Language (EFL) context remains scarce.

Listening does not only require the recognition of words and grammar, but it also involves the complex and dynamic processes of interpreting cues of prosodic and phonological elements such as stress, rhythm, and intonation (Field, 2008). Accordingly, the use of AI-generated speech in high-stake testing settings might highlight critical but necessary questions pertaining to authenticity, quality, and addressing the real-world demands of assessment. Mobile applications and pronunciation tools such as Duolingo and ELSA Speak (Shinde, 2024), have pioneered AI-generated speech in language learning by employing AI-generated voices and exposing language learners to a wide variety of pronunciation models.

The literature accentuated how learners' perception in AI speech could be helpful for listening skills development, specifically when human-like qualities are maintained (Moussalli & Cardoso, 2020, this is likely in accordance with a sense of detachment or reduced engagement when the voice sounds excessively robotic with little to no coherence in emotional depth. These findings underscore the reputation of voice quality and human-comparable authenticity in educational contexts which suggests delicate requirements in propelling the potential and impact of AI speech in testing environments.

As the advancement of synthetic voice generation such as the text-to-speech (TTS) technology via model like 'WaveNet' has suggestively improved the human-comparable naturalness (Shen et al., 2018), along with 'Deep Voice 3' featured as the state-of-the-art neural speech synthesis systems with faster machine-learning magnitude Ping et al., 2017), more exploration is needed to further inform the use of these synthetic voices that aimed to resemble natural spoken language via synthesizing phoneme sequences into waveforms. In contrast to earlier systems, neural TTS has increased the precision from which a more accurate replication of prosodic features, from which these elements are critical for listening comprehension. Even so, several

e-ISBN: 978-629-98082-2-0

limitations in this earlier generation of synthetic voices may influence how listeners process the audio input and ultimately impact listening comprehension.

The gap in literature accentuated the central underlying challenge in using AI-generated speech for listening assessments, where it is indicated towards the importance of its perceived human-likeness. Unlike written or text-based medium, spoken language carries pragmatic, prosodic, and paralinguistic features that influence the comprehension and judgment of listeners (Munro & Derwing, 1995). When emulation of these features is not beyond optimal execution, it might risk the validity and fairness of the assessment. Such predicaments might affect communicative performance in language learning especially since most classrooms today emphasize communicative competence like the CEFR-aligned curriculum whereby assessments are meant to complement them. Accordingly, the validation of AI-powered speech tools demands evaluation by experienced professionals in communicative language curriculum and assessment such as the CEFR. Despite growing interest in AI applications for language testing, empirical research on the use of AI-generated speech in EFL listening tests remains limited.

1.1 Speech & Comprehension Features

The crucial suprasegmental cues of as word stress, rhythm, and intonation are identified features that are widely recognised in facilitating speech segmentation, listener parsing, and the conveyance of speaker intent (Cutler, Dahan, & Van Donselaar, 1997; Field, 2008). As it has been established that stress and timing variations would lead lexical activations and syntactic parsing (Cutler, Dahan, & Van Donselaar, 1997), another dispute by Field (2008) exclaimed how these prosodic features are technically essential for learners to successfully decipher meaning from connected speech. At the segmental level, accuracy in terms of pronunciation along with appropriate enunciation ensures phonemes and word forms remain distinct, simultaneously reducing the taxing effort for listeners to decipher and propelled better comprehension (Celce-Murcia, Brinton, & Goodwin, 2010; Roach, 2009). The importance of articulatory clarity was emphasized by Celce-Murcia et al. (2010) from which classroom-tested techniques were accentuated, while Roach (2009) entails how precision in segmental production strengthens intelligible spoken English. Lastly, the comprehension component of intelligibility was defined as the degree in which spoken language is understood without excessive strain from the listener, in which the combination of prosodic and segmental dimensions is made into a single measure of spoken-language efficacy (Munro & Derwing, 1995; Derwing & Munro, 2005). By considering these six interrelated constructs of speech features and comprehension, this study employed a comprehensive framework for evaluating the human-likeness and real classroom assessment suitability of AI-generated speech in EFL listening assessment.

e-ISBN: 978-629-98082-2-0

This study aimed to contribute to this emergent need by investigating the use of AI-powered speech among vis-à-vis the identified features as construct of judgement by English language assessment experts. This study specifically explored the judgments of expert informants on the human-likeness compatibility of AI-generated speech across six identified dimensions of speech and comprehension features, which is word stress, rhythm, intonation, enunciation, pronunciation accuracy, and intelligibility. These identified features made as evaluation constructs were not only essential for evaluating human-comparable spoken communication but also intended to echo the descriptors of communicative needs within CEFR-based listening proficiency levels (North & Piccardo, 2016).

2. Methods

To investigate these issues, this study employed a two-phase methodology. The methodology incorporates both quantitative and qualitative data to propel a comprehensive evaluation on the features of the AI-powered speech across the aforementioned dimensions of speech and comprehension components. Evaluation criteria were identified based past studies in which the identification of relevant features that influence efficacy of listening assessment in tandem with established research on speech perception (Munro & Derwing, 1995; Field, 2008). A set of structured items was developed and refined based on the identifiable criteria in literature. The primary focus was to gauge expert evaluation and reach a level of expert consensus in tandem with the proximity of how the AI-powered speech compares to human-likeness in the context of listening test for the English language.

2.1 Mixed-Method study

In the first phase, the Fuzzy Delphi Method was used to quantify the expert consensus among seven expert informants who are certified CEFR Master Trainers on the human-likeness of the AI-powered speech across the identified features. FDM is an effective tool for managing uncertainty and subjectivity in expert judgments, making it suitable for research involving perceptual features of language (Ishikawa et al., 1993). The fuzzy values derived from expert responses were defuzzified to calculate the *threshold value (a)*, *consensus percentage*, and *average expert agreement*, which served as indicators for acceptance or rejection of each evaluated dimension. Following that, a phase of focused group interviews and individual interviews provide qualitative insights to further excavate understanding of the expert informants' judgment and anecdotal nuances behind their rating decisions, all while rationalizing the ratings with their overall agreement consensus. The interviews were audio-recorded, transcribed verbatim, and subjected to thematic analysis using Braun and Clarke's (2006) framework. Coding focused on recurring themes to complements the statistical analysis of FDM by providing deeper insights how expert judgments were formed and what influence

e-ISBN: 978-629-98082-2-0

their rating dispersion or agreeable cumulative ratings. This mixed method of data collection aimed to explore the nuances behind agreement levels, gauge perceived strengths and limitations of the AI-powered speech and excavating further recommendations for a more optimal AI-driven integration in future cycles.

The interviews were semi-structured and steered by the six (6) features and emerging themes from the first phase, in areas of divergent consensus. This approach allows for nuanced expert's input while accommodating any room of ambiguity in linguistic judgments (Ishikawa et al., 1993). This methodological triangulation enhanced the validity of the study's conclusions (Creswell & Clark, 2018). By merging the Fuzzy Delphi Method and qualitative interviews, this two-phase design aimed to increase both the reliability and depth of findings. The methodology not only quantifies expert consensus but the thematic analysis for the qualitative data also opens interpretative space to explore assessment-related implications of integrating AI-powered speech in real-world testing contexts. Conclusively, this study offers a robust framework for evaluating the viability of using AI-powered speech as test stimuli in language assessment, particularly for listening assessment in EFL context, reinforced by findings from experts of CEFR-aligned assessments.

The research is framed by two main research questions: (1) To what extent do the expert informants (language assessment experts) consider the AI-generated speech as human-like in terms of *word stress*, *rhythm*, *intonation*, *enunciation*, *accuracy (pronunciation accuracy)*, and *intelligibility*, along with (2) Which features from the AI-powered speech are perceived as most or least naturally emulated. In addressing these research questions, this study contributes to the validation of AI-driven tools for language testing and simultaneously informs the ongoing conversation on human-machine interaction in educational landscape, particularly for assessment contexts. The findings were expected to support the principled integration of AI technologies in language learning in tandem with safeguarding the alignment of such tools with real-world communication demands based on key communicative needs in CEFR-based classrooms.

2.2 Audio Text

Experts were asked to evaluate the degree of human-likeness of the AI-powered speech stimuli using a 10-point linguistic Likert scale. The AI-powered speech of a listening text stimuli was generated by a commercial AI text-to-speech engine configured for a near neutral accent. Readability Indexes were calculated to justify the readability of the audio for the audience in which the listening text is analysed with the Flesh Reading Ease, Flesh Kincaid Grade level along with the calculation of Gunning-Fox index as shown in Table 1 below.

e-ISBN: 978-629-98082-2-0

Table 1: Readability Indexes for the Listening Audio Text

Text	*Flesh reading Ease*	*Flesh Kincaid Grade level*	*Gunning-Fox index*	*Lexical Sophistication (CEFR)*
Audio Text 1	71.43	4.92	7.02	B2 (52%)

Source: Results of readability indexes calculated based on the transcription of the audio text selected for this study

The listening text scored a Flesch Reading Ease of 71.43, indicating that it is fairly easy to understand for most readers. The Flesch-Kincaid Grade Level of 4.92 suggests that the content is suitable for learners around the 5th-grade level, aligning with B1–B2 in the CEFR Global scale's expectations for communicative English. In addition, the Gunning Fog Index of 7.02 indicates the text would be easily understood by individuals with at least 7 years of formal education, in tandem with suitability of utilizing this stimuli for entry-level EFL university learners. Readings on the overall lexical sophistication of the audio text was B2 plus for most of the text by 55%, for which the text considered suitable for any input text found in CEFR-aligned listening tasks for entry-level EFL university learners.

2.3 Expert Informants

In this study, a total of seven (7) expert informants were selected based on their specialized knowledge and shared professional background. Although it is highly recommended to obtain a panel size of between 10 to 50 (Jones & Twiss, 1978), or from 10 to 15 (Adler & Ziglio, 1996), the smaller number used in this study was employed due to the high degree of criterion among participants, expert homogeneity is high as all of them are selected based on these criteria:

1. A CEFR Master Trainer certified by Cambridge for CEFR-aligned curriculum and assessment in Malaysia.
2. A state-level coach instructor with experience in training teachers for CEFR-aligned curriculum and assessment in Malaysia.
3. An English language teacher with more than 5 years of experience in schools.
4. A bachelor's degree holder or a master's degree holder in Education (Teaching English as a Second Language).

Based on the established criteria, 7 expert informants have been shortlisted. The experts have been identified from 5 schools, and 2 government-linked bodies that have the qualification (criteria 4), CEFR-aligned curriculum and assessment certification (criteria 1), extensive experience in teaching CEFR-aligned curriculum (criteria 3) along with experience of training teachers and implementing the CEFR-aligned curriculum and assessment across the country (criteria 2). In tandem with the standard of procedure for data collection, informed consent was obtained via consent forms prior to data collection.

e-ISBN: 978-629-98082-2-0

2.4 Data Analysis Questionnaire

To evaluate the AI-powered speech, a 10-point Likert scale was adopted to collect expert informants' judgments. Although 5-point and 7-point scales are more commonly seen in language testing literature, a 10-point scale was intentionally selected to achieve greater granularity in measuring the nuanced expert perceptions, which are considered particularly vital in assessing multi-dimensional constructs that combines both speech and comprehension features of spoken language such the six features in this study. This is supported by prior research on the Fuzzy Delphi Method (FDM), which recommends broader scales for more precise defuzzification and improved sensitivity of measurement (Jamil et al., 2016; Mustapha et al., 2019). Since the expert informants were selected based on a few criteria that encapsulates experience and high familiarity in communicative-aligned speech assessment, the cognitive taxation of 10-point scale was considered manageable and even beneficial for further distinguishment between features of evaluations. Thus, the 10-point scale not only aligns with established practices in FDM-based educational studies but also enhances the validity and resolution of the consensus measurement process.

The data analysis was carried out in accordingly, with the experts' viewpoints thoroughly examined using Microsoft Excel, as highlighted by Ramlie et al. (2014), and Mohd Jamil et al. (2013). As the Triangular Fuzzy Number (TFN) and the Defuzzification Process are vital parts of the Fuzzy Delphi Technique, each expert informant's judgment on a given criterion of identified feature was represented as a triangular fuzzy number. This number is defined by a lower bound, a most likely (modal) value, and an upper bound. To combine the n experts' assessments, all of the lower bounds needed to be averaged to form an overall lower score, the same should be done similarly for the modal and upper bounds. The average of those three aggregated scores should be taken to produce a single crisp value for use in subsequent analysis (Klir & Yuan, 1995). Defuzzification calculation converts each aggregated triangular fuzzy number into a single crisp rating score. After combining all experts informants' lower, modal, and upper bounds by averaging them separately, the final score is calculated by taking the mean of those three aggregated numbers. In other words, the "defuzzified" score is simply the average of the overall lower bound, the overall most-likely value, and the overall upper bound, providing a straightforward, single metric for further analysis.

The Triangular Fuzzy Number also involves assessing the percentage of expert agreement as the next condition as established by (Chu & Hwang, 2008; Murray & Hammons, 1995) where consensus is attained and acknowledged when the agreement among the expert group exceeds 75%. This signifies the acceptable range of which the percentage of evaluations in individual features in tandem with the traditional Delphi technique. The Defuzzification is implemented by converting fuzzy numbers into specific scores, from which a ranking of the identified speech

e-ISBN: 978-629-98082-2-0

and comprehension features of the AI-powered speech will be tabulated from highest to lowest. The Defuzzification Process comprises of obtaining the fuzzy score value (A), grounded on an α-cut value of 0.5 where the item being measured is accepted if the fuzzy score (A) is 0.5 or greater while this is in opposition with the case if the item is rejected if it is below 0.5. The fuzzy (A) score value is calculated using the formula as shown in (Eq. 1).

$$A^* = \frac{l + m + u}{3}$$

(1)

Rank of the items or features that were evaluated will also be tabulated in which the final verdict of whether the expert consensus is achieved following the determined circumstances of each item to obtain threshold value; α-cut of ≤ 0.5 and a minimum of 70% expert agreement. In assessing the consensus, the α-cut threshold (Zadeh, 1965) quantifies the average width of the TFNs which cuts to the 0.5 level and retained only those values whose confidence was at least 0.5. Any value below this cutoff was considered as lacking consensus and therefore excluded from the final verdict on acceptable expert consensus, therefore informed the conclusions of the findings. In the Fuzzy Delphi Method, the α-cut threshold of 0.50 is affixed to the mathematical definition of triangular fuzzy numbers, representing the midpoint between full disagreement (0) and full agreement (1). Chen and Lin (2002) first articulated this by specifying that a defuzzified score (A) must meet or exceed 0.50 to signify genuine expert consensus. The use of the benchmark of the α-cut for the threshold of acceptable expert consensus is further confirmed by Mohd Ridhuan et al. (2013) whose study further clarified that using α-cut = 0.50 balances the risk of false positives (accepting marginal items) against false negatives (rejecting borderline items). The formula of the α-cut for the threshold is shown in (Eq. 2) below.

$$A\alpha = \{x \mid \mu A(x) \geq \alpha\} = [\ell + \alpha(m - \ell), u - \alpha(u - m)]$$

(2)

Despite varied perspectives on defining consensus, justification of the decision for the expert consensus to be acceptable to the study must be made (Jorm, 2015; Hall, et. al., 2018). As such, the Traditional Delphi methodology often applies the majority agreement threshold of 70% to define consensus, a benchmark especially relevant in studies with small expert panels such as the study by Young et al. (2022) whose predetermined consensus by at least 70% to be essential by experts, although it was done in a three-cycle datasets. Other studies also considered consensus of an item with 70% of experts responding the agreeable range of their item (Romero-Collado 2021; Humphrey et al, 2017) Thus, for the purpose of this study with only seven experts, adopting a 70% agreement benchmark maintains methodological precision by capturing true majority views while agreeing to individual divergence.

e-ISBN: 978-629-98082-2-0

3. Results and Discussion

3.1 Quantitative Findings: Fuzzy Delphi Analysis on the AI-powered speech

The quantitative phase of this study applied the Fuzzy Delphi Method (FDM) to determine expert consensus on six features of AI-generated speech for listening test use stated in Table 2. Table 2 shows the identified features of speech and comprehension in spoken English that were made as the construct for the items that were measured by the expert informants. The features were word stress, rhythm, intonation, enunciation, pronunciation accuracy, and intelligibility.

Table 2: Items for the Evaluation of AI-powered speech for Listening Test

Feature	*Item*
F1	Word Stress
F2	Rhythm
F3	Intonation
F4	Enunciation (Clarity of Articulation)
F5	Pronunciation Accuracy
F6	Intelligibility

Source: Features identified for measurement based on past studies

Seven (7) certified and experienced experts informants selected based on criterion sampling provided their ratings on a 10-point Likert scale. The data collected were transformed into triangular fuzzy numbers, followed by defuzzification and threshold analysis. Items from the construct identified in Table 2 were measured using a 10-point Likert scale, from zero (0) which is labelled as extremely low to ten (10), which is labelled as extremely high. Results of the survey were analysed, and the expert consensus was calculated. A threshold value (α-cut) of ≤ 0.5 and a minimum of 70% expert agreement were used as benchmarks of consensus based on the aforementioned determined threshold. The following table 2 summarises the findings from the FDM analysis:

Table 3: Findings of FDM Expert Consensus

Item	*TFN (average minimum, best judgement, maximum)*	*Defuzzified Value (A)*	*Ranking*	*Threshold, α-cut ≤ 0.5*	*Percentage of Expert Agreement, %*	*Expert Consensus*
F1 Word Stress	*(7.14, 8.43, 9.86)*	*8.452*	*4*	*0.429*	*71%*	*Accepted*
F2 Rhythm	*(4.29, 4.57, 6.29)*	*4.81*	*6*	*0.095*	*86%*	*Not Accepted*

e-ISBN: 978-629-98082-2-0

F3 Intonation	*(4.86, 6.57, 8.0)*	*6.524*	*5*	*0.536*	*57%*	*Not Accepted*
F4 Enunciation	*(8.14, 8.71, 10.0)*	*8.833*	*3*	*0.286*	*100%*	*Accepted*
F5 Accuracy (Pronunciation)	*(8.29, 9.43, 10.0)*	*9.333*	*1*	*0.714*	*57%*	*Not Accepted*
F6 Intelligibility	*(7.43, 9.0, 9.71)*	*8.857*	*2*	*0.738*	*29%*	*Not Accepted*

Source: Quantitative data of expert informants' ratings

Table 3 summarises the findings of the Fuzzy Delphi Method (analysis) for the extent of how the expert informants consider the AI-generated speech as human-like based on the identified features of speech and comprehension in spoken English. The Triangular Fuzzy Numbers were calculated, in which the average minimum rating, average best judgment rating and average maximum rating were calculated and tabulated. The Defuzzified numbers (A) were calculated and ranked from highest to the lowest, followed by the α-cut threshold value and the percentage of expert agreement among the expert informants for each feature. The table ended with an expert consensus based on the pre-determined benchmark.

The expert informants' consensus was evaluated using two criteria: α-cut threshold ≤ 0.5 and at least 70% agreement among the seven experts. Based on the α-cut calculations, features were evaluated for expert consensus. The six features of speech and comprehension were evaluated and two features (word stress and enunciation), were identified to have met this threshold. The features of word stress and enunciation were indicated to obtain confident and narrow value of agreement among the expert panel. The other four features such as Rhythm, Intonation, Accuracy and Intelligibility did not adhere to the determined benchmark of consensus via the α-cut threshold, illustrating a higher degree of variability or less certainty in expert judgements among the expert informants. Although the threshold value of (d) was also calculated to observe the inter-rater dispersion, it was not used as the basis for consensus determination in this study as the selection of this determined criteria reflects a methodological emphasis on individual confidence as taken by the calculation of the α-cut, which is intended for FDM applications with smaller scale of expert informants.

Pronunciation accuracy received the highest defuzzified value (A = 9.333, rank 1), indicating it was the top-rated feature in terms of perceived human-likeness. However, this feature failed to reach consensus, with only 57% of experts in agreement and an α-cut threshold of 0.714 (exceeding the 0.5 cutoff). Intelligibility was the second-highest rated feature (A = 8.857, rank 2), but it likewise did not achieve consensus due to a low agreement rate of 29% and a relatively

e-ISBN: 978-629-98082-2-0

high α-cut value of 0.738. In contrast, Enunciation (clarity and articulation) was among the top-rated features (A = 8.833, rank 3) and successfully met the consensus requirements. All experts (100% agreement) concurred on enunciation, and its α-cut threshold was 0.286, well below the 0.5 criterion, indicating a strong consensus for this feature. Word stress (A = 8.452, rank 4) was the other feature that achieved consensus, with 71% of the panel in agreement and an α-cut value of 0.429; both metrics satisfy the predetermined thresholds.

The remaining lower-rated features did not reach consensus. Intonation had a moderate defuzzified value (A = 6.524, rank 5) with 57% expert agreement and an α-cut threshold of 0.536, failing to meet the 70%/0.5 consensus benchmarks. Rhythm was the lowest rated feature (A = 4.810, rank 6), reflecting the weakest human-like performance among the six. Despite its low rating, rhythm showed a relatively high percentage of expert agreement (86%) and the smallest α-cut threshold value (0.095) of all features, indicating a narrow spread in the experts' ratings. Nevertheless, rhythm was not accepted as reaching consensus in the final analysis, as it did not fulfill all the required criteria for consensus acceptance. In summary, only word stress and enunciation emerged with clear expert consensus (α-cut $\leq$ 0.5 and $\geq$ 70% agreement), whereas pronunciation accuracy, intelligibility, intonation, and rhythm did not meet one or both of these quantitative consensus thresholds (see Table 2).

3.2 Qualitative Findings: Interview with Expert informants on the AI-powered speech

The qualitative phase of this study involved a series of interview sessions. Factors such as time and proximity constraints influenced the decision of gauging qualitative data through a mix of focused group discussions and individual interviews with all of the experts informants from Respondent 1 (R1) to Respondent 7 (R7). The qualitative data is aimed at further investigating the understanding of the AI-generated speech tool's performance via a follow-up insight on their decision-making and gaining consensus among the expert informants. These insights are focused on the six speech features: word stress, rhythm, intonation, enunciation, accuracy (or pronunciation accuracy), and intelligibility. Thematic analysis was employed by identifying patterns across the responses of these expert informants while simultaneously aiming to connect with the quantitative findings from the Fuzzy Delphi Method.

All of the interviews were recorded and transcribed, followed by iterative review for familiarisation, which is recommended by Braun and Clarke (2006) as this stage allows the researcher to familiarise themselves with the complexity and the extent of the content, in which transcription is a significant segment of data analysis in qualitative research. A thorough thematic analysis was implemented following Braun and Clarke's (2006) method. A deductive coding was employed from which the pre-established combination of speech and comprehension features as the construct of evaluation steered the conversation but emergent

e-ISBN: 978-629-98082-2-0

subthemes were investigated and discussed, such as perceived artificiality, speaker comparison, audience compatibility and suitability of employing the AI-powered speech in real-life listening assessment. In systematically compare responses across the identified features, a thematic matrix was used, which aligned with the data display principles of Miles, Huberman, and Saldaña (2014), which allowed for a structured comparison of codes and insights across respondents and themes that promote the processes of pattern recognition, explanation building, and conclusion drawing. This method enabled visualization of features that were highly rated such as the Enunciation and Pronunciation Accuracy feature, while simultaneously highlighting the problematic parts which were rhythm and intonation, specifically pertaining to naturalness.

Findings from the thematic analysis were later integrated with the FDM results using a joint display matrix, which highlighted both convergence (e.g., expert consensus on intelligibility) and divergence (e.g., varying perceptions of rhythm based on gendered voice pacing). This methodological triangulation enhanced the validity of the study's conclusions (Creswell & Plano Clark, 2018).

Table 4: Findings from Interview Based on FDM Features of AI-powered Speech

Features from FDM & Emerging Themes	*Summary Insight*
Word Stress	Generally accurate but less salient to some (less than half) at the start of the conversation.
Rhythm	Perceived as mechanical or "rehearsed" pacing in some parts
Intonation	Varied opinions; clear in some areas but lacking emotional nuance in others
Enunciation	Generally praised for clarity
Pronunciation Accuracy	Rated highly; some accents considered intelligible and accessible for the common ear
Intelligibility	Overall clear and comprehensible, usable for classroom contexts

Source: Qualitative thematic matrix from the interview transcription

The Qualitative insights were prevalent and further discussed the features identified and measured from the FDM rating and translated into a deeper discussion that illustrated the divergence and consensus attained by the previous phase of data collection. Summaries of insights by feature were tabulated in Table 4 above. All of the expert informants consistently praised the clarity and correctness of pronunciation, from which words were articulated accurately and without noticeable accent interference which some noted how it was "near-perfect" (R5). However, variability in scoring intensity as some expert informants avoided giving the highest rate of 10 led to dispersion in ratings. In terms of voice projection, the AI-

e-ISBN: 978-629-98082-2-0

powered speech was generally regarded by a majority of the expert informants as clear and well-articulated. Raters could understand every word, although only one expert respondent noted that the male speaker's enunciation was slightly clearer than the female (R6). Even so, this minor discrepancy contributed to the scoring variation.

Expert informants agreed that the word stress displayed by the AI-powered speech was accurate by most of them. Even so, some (R4 & R6) observed that stress was not prominent at the beginning of the audio but became clearer over time. This variation in perception influenced the scores across respondents. As juxtaposed to that, general opinions on intonation were mixed. While some found the pitch variation adequate (R4 & R6), others described the timing of the back and forth between the speakers sounded as rehearsed (R7), a little robotic (R2), and at times lacking emotional authenticity (R1, & R5). The inconsistency in how well emotional cues were conveyed led to moderate ratings and low agreement.

The most problematic feature identified and expressed by all of the expert informants were rhythm. Experts repeatedly noted that the pacing of speech felt unnatural at times, particularly for the female speaker (R4 & R6). The flow of the conversation was perceived as choppy (R3) or overly mechanical (R2, R7), which significantly affected its perceived human-likeness. Even so, some of the expert informants (R1, R2 & R3) confirmed that noted how the rhythm felt technically off but only at minimal part of the speech and not at the totality of the whole audio. Lastly, intelligibility emerged as a strong theme in the qualitative data. All experts indicated that the audio was comprehensible and that students would be able to follow the message. However, the clarity was somewhat affected by rhythm and pacing as noted in the earlier notes on the problematic features of rhythm and intonation. Insights from the interview noted that initially, some of the expert informants had to exert a little degree of effort to adapt to the pacing for further comprehension during the initial stage of listening.

Table 5: Other Emerging Themes Based on the Interview

Features from FDM & Emerging Themes	*Summary Insight*
Exemplar Language Command	Highlighted for exceptional accuracy and suitable to be a model exemplar for classroom reference
Naturalness & Engagement	Noted gaps between scripted delivery and spontaneous human-like conversation, although at minimal level
Classroom Usability	Collectively accepted to be suitable for listening assessments in tertiary and upper-secondary learners with minor room for improvement

Source: Qualitative thematic matrix from the interview transcription

e-ISBN: 978-629-98082-2-0

Table 5 shows other emerging themes that evades from the distinct speech and comprehension features were noted in the interview, as all of the expert informants accentuated how the AI-powered speech is usable for classroom use and could be a great sample for pronunciation and overall general language command. Even so, the unnatural pacing of the exchange between speakers in the audio may made it look rehearsed and unnatural to authentic exchange between speakers. Some degree of emotional authenticity were also explicated by the expert informants that it needed to be polished in order to resemble human-comparable traits that is vital to distinguish speakers underlying intention and overall tone of the conversation which would support listening comprehension. Even so, the AI-powered received praises and were claimed to be exceptionally usable for classroom use. Some even noted how it is generally ready for tertiary to upper-secondary school learners. Overall, the AI-generated speech was generally perceived as technically accurate and intelligible, with strong enunciation and acceptable word stress. Even so, the aim to achieve human-like naturalness in rhythm and intonation remains an area for improvement. While the AI-powered speech may be viable for use in language assessments, further refinement is needed to enhance the perception of natural, human-like delivery to address real-world demand of listening assessments.

4. Discussion

This study examined six speech and comprehension features of AI-powered speech, which are Word Stress, Rhythm, Intonation, Enunciation, Pronunciation Accuracy, and Intelligibility. By integrating quantitative consensus measures from the Fuzzy Delphi Method (FDM) with qualitative insights from the interviews with expert informants, the culmination of the quantitative and qualitative data sources illuminates where the AI-powered speech excels and where rooms for improvements were addressed to attain the human-comparable goals which will suit real-world listening assessment demands.

Quantitative calculation on the Fuzzy score of Enunciation and Word Stress was shown as the only features to satisfy the dual consensus criteria established for this study (in which α-cut $\leq$ 0.50 and $\geq$ 70% agreement; see Table 3). Enunciation achieved undivided agreement by 100% and a low α-cut of 0.286, while word stress surpassed the agreement threshold by 71% with α-cut of 0.429. Qualitative insights from the experts unanimously praised these two features whereby some of the expert informants agreed that “each word was pronounced with care and the clarity made it easy to understand” (R5) and that “everything was on point” with respect to stress placement (R7). This convergence indicates that the AI-powered speech’s segmental and syllabic control reliably meets expert expectations for clear, intelligible test stimuli.

e-ISBN: 978-629-98082-2-0

In contrast, Pronunciation Accuracy and Intelligibility ranked first and second in defuzzified value (A) of 9.333 and 8.857, respectively but did not achieve quantitative expert consensus agreement, owing to α-cuts of 0.714 and 0.738 and agreement rates of 57% and 29%. Even so, in the interviews expert informants elucidated that the Pronunciation feature was "near perfect" (R5) and Intelligibility as "quite straightforward" (R1), with several noting how they were able to follow the message and would "not need to guess the message" (R3). These qualitative confirmations suggest that although some experts' fuzzy ratings were varied enough to prevent acceptable consensus with regards to the benchmark established by this study, the overall perception of accurate pronunciation and clear comprehension remains positive and highly rated. The divergence appears to stem from the strictness of the established FDM thresholds when applied to a small expert panel rather than any fundamental deficiency in the AI-powered speech.

Finally, the prosodic features of Intonation and Rhythm were unsuccessful to satisfy consensus benchmarks with results of α-cut ≥ 0.536 and 0.095, followed by consensus agreement of 57% and 86%, respectively, in which these features were also the lowest-ranked features. Interview data mirror this mixed reception with expert informants describing Rhythm as "a bit choppy" (R3) or "rehearsed" (R1), whereas only a minority of them found it "smooth and consistent" (R5). Intonation similarly divided the expert informants, in which some found pitch variation as "lively and expressive" (R5), while the majority judged it "not 100% there" and overly scripted (R7). Conclusively, these findings indicate that although the AI-powered speech has the proximity to human-comparable prosody, it is yet to attain the full dynamism and natural timing of conversational human-like speech.

The strong alignment between quantitative and qualitative measures for Enunciation and Word Stress supports the immediate need of AI-generated speech in practical contexts where clarity and correct syllabic emphasis are emphasised. Even so, the lack of consensus on Rhythm and Intonation, which gravitates to authentic communicative competence would suggests caution when designing tasks or assessments that targets processing natural prosody such as measuring inference towards speaker's attitude or tonal nuances. AI-driven test designers should prioritise further refinement in prosodic modelling, such as integrating authentic pitch contours, varied interactional exchange timing or even room for filler words or sounds to emulate authentic human-like conversation in tandem with meeting the real-world communicative competence demands in listening assessments.

e-ISBN: 978-629-98082-2-0

5. Conclusion

By triangulating FDM consensus metrics with expert interview data, this study accentuated a nuanced judgement of AI-powered speech for tertiary level listening test for EFL learners. Enunciation and word stress are both quantifiably and experientially robust, whereas prosodic elements require refinement. Though Pronunciation accuracy and intelligibility were highly rated, the rating dispersion exhibited was enough to fall short of general agreeable consensus between the expert informants. This might be derived from the limitations of methodology rather than data deficiency. Overall, the convergent evidence confirms that the AI-powered speech was able to deliver clear, well-articulated speech for EFL listening test contexts, while highlighting further refinement of human-comparable prosody as the next frontier for human-like speech synthesis in educational assessment.

6. Limitations and Future Research

This study only reported the first cycle of data collection, from which another cycle is expected to further inform the precision of each speech and comprehension feature of AI-powered speech. In tandem with the small sample size of (n = 7), the FDM consensus thresholds may have been sensitive to individual rating variability, particularly for pronunciation and intelligibility, despite their high ratings. Future cycles will aim to expand the expert sample to increase the quantitative robustness or adjust the agreement benchmarks for scalable studies. While qualitative data interviews provided nuanced perspectives, expert informants only assessed a single AI-powered speech sample. Further research would examine wider range of voices models and content (audio text) readability to generalise findings.

7. Acknowledgement

This research was supported by the Centre for Shaping Advanced & Professional Education (UKMShape) and Dr. Ehsan Mohd Said from the Faculty of Education at Universiti Kebangsaan Malaysia.

References

Adler, M., & Ziglio, E. (1996). *Gazing into the oracle: The Delphi method and its application to social policy and public health*. Jessica Kingsley Publishers.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chang, L., & Cutumisu, M. (2021). Integrating deep learning into an automated feedback

e-ISBN: 978-629-98082-2-0

generation system for automated essay scoring. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)* (pp. 573–579).

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide*. Cambridge University Press.

Chen, C.-H., & Lin, C.-T. (2002). Fuzzy Delphi method for evaluating educational program elements. *Asian Journal of University Education, 17*(1), 298–310.

Chu, H. C., & Hwang, G. J. (2008). A Delphi-based approach to developing expert systems with the cooperation of multiple experts. *Expert Systems with Applications, 34*(4), 2826–2840. https://doi.org/10.1016/j.eswa.2007.05.044

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume*. Council of Europe Publishing.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*(2), 141–201. https://doi.org/10.1177/002383099704000203

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397. Retrieved from https://www.jstor.org/stable/3588486

Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.

Hall, D. A., Smith, H., Heffernan, E., & Fackrell, K. (2018). Recruiting and retaining participants in e-Delphi surveys for core outcome set development: Evaluating the COMiT'ID study. *PLoS One, 13*, e0201378. https://doi.org/10.1371/journal.pone.0201378

Humphrey-Murto, S., Varpio, L., Gonsalves, C., & Wood, T. J. (2017). Using consensus group methods such as Delphi and Nominal Group in medical education research. *Medical Teacher, 39*, 14–19. https://doi.org/10.1080/0142159X.2017.1245856

Ishikawa, A., Amagasa, M., Shiga, T., Tomizawa, G., Tatsuta, R., & Mieno, H. (1993). The max-min Delphi method and fuzzy Delphi method via fuzzy integration. *Fuzzy Sets and Systems, 55*(3), 241–253. https://doi.org/10.1016/0165-0114(93)90251-C

Jorm, A. F. (2015). Using the Delphi expert consensus method in mental health research. *Australian & New Zealand Journal of Psychiatry, 49*, 887–897. https://doi.org/10.1177/0004867415600891

Jamil, M. F. M., Alias, M., & Azmi, I. A. G. (2016). Fuzzy Delphi method for decision-making in technical and vocational education. *Malaysian Journal of Society and Space, 12*(2), 129–137.

Jones, H., & Twiss, B. C. (1978). *Forecasting technology for planning decisions*. Macmillan.

Klir, G. J., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: Theory and applications*. Prentice

e-ISBN: 978-629-98082-2-0

Hall.

Kuo, Y. F., & Chen, P. C. (2008). Constructing performance appraisal indicators for mobility of the service industries using Fuzzy Delphi Method. *Expert Systems with Applications, 35*(4), 1930–1939. https://doi.org/10.1016/j.eswa.2007.08.118

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

Mohd Jamil, M. F., Abas, N. H., & Ariffin, N. H. (2020). Application of the Fuzzy Delphi Method (FDM) in determining expert consensus for TVET curriculum requirements. *Journal of Technical Education and Training (JTET), 12*(2), 117–124. https://doi.org/10.30880/jtet.2020.12.02.012

Mohd Jamil, M. R., Mat Noh, N., Sulaiman, N. D., Sham, R., & Mohd Nasir, B. (2013). Fuzzy Delphi technique as a tool for eliciting expert opinion: A case study on the construction of a competency framework for the Malaysian construction industry. *Jurnal Teknologi, 65*(1), 1–10. https://doi.org/10.11113/jt.v65.1958

Mohd Ridhuan, M., Jamil, A. F. M., et al. (2013). The application of the Fuzzy Delphi technique to parental involvement in preschool education. *Journal of Education and Practice, 4*(5), 88–102.

Moussalli, S., & Cardoso, W. (2020). Text-to-speech synthesizers in the EFL classroom: Students' perceptions and performance. *Computer Assisted Language Learning.*

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Murray, J. W., & Hammons, J. O. (1995). Delphi: A versatile methodology for conducting qualitative research. *The Review of Higher Education, 18*(4), 423–436. https://doi.org/10.1353/rhe.1995.0008

Mustapha, R., Adnan, A. H. M., & Ab Rahman, N. S. F. (2019). Application of Fuzzy Delphi Method (FDM) in education: A review. *International Journal of Academic Research in Business and Social Sciences, 9*(1), 144–156. https://doi.org/10.6007/IJARBSS/v9-i1/5361

North, B., & Piccardo, E. (2016). Developing illustrative descriptors of aspects of mediation for the Common European Framework of Reference (CEFR). Council of Europe.

Ping, W., Peng, K., Gibiansky, A., Arik, S. Ö., Kannan, A., Narang, S., … & Miller, J. (2017). Deep Voice 3: Scaling text-to-speech with convolutional sequence learning.

Roach, P. (2009). *English phonetics and phonology: A practical course* (4th ed.). Cambridge University Press.

Romero-Collado, A. (2021). Essential elements to elaborate a study with the (e)Delphi method. *Enfermería Intensiva, 32*(2), 100–104. https://doi.org/10.1016/j.enfie.2020.09.003

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., … & Wu, Y. (2018). Natural

e-ISBN: 978-629-98082-2-0

TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779–4783). https://doi.org/10.1109/ICASSP.2018.8461368

Shinde, S. M. (2024). Leveraging AI for English language learning: A comparative analysis of Duolingo, Babbel, and ELSA Speak. *Language in India, 24*(10), 26–35. Retrieved from languageinindia.com

Susanti, M. N. I., Ramadhan, A., & Warnars, H. L. H. S. (2023). Automatic essay exam scoring system: A systematic literature review. *Procedia Computer Science, 216*, 531–538. https://doi.org/10.1016/j.procs.2022.12.166

Young, A. M., Chung, H., Chaplain, A., Lowe, J. R., STARS Rehabilitation Development Group, & Wallace, S. J. (2022). Development of a minimum dataset for subacute rehabilitation: A three-round e-Delphi consensus study. *BMJ Open, 12*(3), e058725. https://doi.org/10.1136/bmjopen-2021-058725

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

e-ISBN: 978-629-98082-2-0

P049

Investigating Multiple-Choice Test Items Designed for Specific Reading Constructs: Insight from Item Writing

***Pham Ngoc Bao Tram[1], Zeng Yijing[2], Khairi Fakhri Fazil[3], Nur Ehsan Mohd Said[4]**

[1-3]Centre for Shaping Advanced and Professional Education (UKMSHAPE), Universiti Kebangsaan Malaysia, Malaysia
[4]Faculty of Education, Universiti Kebangsaan Malaysia, Malaysia
(Email: [1]phngbaotram@ukm.edu.my, [2]zengyijing@ukm.edu.my, [3]khairifakhri@ukm.edu.my, [4]nurehsan@ukm.edu.my)
**corresponding author: [1]phngbaotram@ukm.edu.my*

Abstract

While writing multiple-choice (MC) test items that accurately target specific reading constructs is challenging, very little research has examined how item writers operationalise reading constructs into MC items with content. This study aims to document the intentions of item writers when constructing MC items that target specific reading constructs, including item features in terms of text explicitness and the rationales behind the key and distractors of each item. The findings reveal that most MC items, despite targeting different constructs, were text-implicit, requiring the integration of information across multiple sentences and bridging inferences to arrive at the correct answer. The analysis also shows variation in the size of relevant text portions and the cognitive demands of different options within the same item. These insights help explain how and why test-takers might be drawn to certain distractors, beyond the assumption of simple miscomprehension. While the study provides detailed documentation of item writers' intentions, there are limitations regarding a small test-taker sample and the absence of interrater coding.

Keywords: L2 reading assessment; item writing; multiple choice test item

1. Introduction

Multiple-choice (MC) format has been the most common item format used in L2 reading assessment due to its advantages in administering and scoring steps (Jeon & Yamashita, 2020). Nevertheless, the challenge lies in designing a good multiple-choice test item that taps into a particular test construct (Jones, 2020; Schedl & Malloy, 2013). Very few studies in language testing have investigated the process of writing test items in general and MC test items specifically (but see, e.g., Green & Hawkey, 2011, 2012; Kim et al., 2010; Ngo, 2020; Rossi, 2021; Salisbury, 2005). The existing studies offer insights into the sequential phases that item

e-ISBN: 978-629-98082-2-0

writers may go through when writing test items and how their approaches may change as their skills develop. However, a micro-level of how item writers operationalise a specific test construct into an item with content (e.g., decide what questions to ask, what distractors to include, and confirm the appropriateness of test items to the provided test specifications) has not been rigorously studied.

Initial judgments by item writers on how a test item may work are important to gauge the skills potentially measured by test tasks (Liu & Read, 2021). However, it has been observed that the reading processes and strategies used by test-takers are not necessarily similar to those that item writers believe their items are eliciting (Alderson, 1990; Rupp et al., 2006). While the documentation of item writers' intentions can allow a direct comparison with test-takers' actual performance to gain insight into the reasons behind the discrepancies (Liu, 2018), it is not uncommon that item writers' intentions when creating items often remain undocumented.

Motivated by these gaps, this study aims to document the intentions item writers have when constructing MC items targeting different reading constructs. The items were constructed for the reading test in the Higher Education English Test (HEET), which aims to assess local and international students' English proficiency for university admission in Malaysia where English is used as a Second/Foreign Language. Specifically, the following research questions are formed:

1.1. What are the features of MC items that target specific reading constructs, in terms of text explicitness?
1.2. What are the rationales behind the key and distractors of MC items that target specific reading constructs?

2. Methods

2.1. Test materials

The reading testlet used in this study was the initial draft created by the original item writers, including a single text and a set of MC reading items. Table 1 below shows a summary of the characteristics of the text.

Table 1: Text characteristics

Topic	Productive debate
Text length	580
Readability (Flesch Reading Ease)	56.21

e-ISBN: 978-629-98082-2-0

2.2. Item descriptions

6 items in the testlet were selected for the analysis in this study. The items were designed to target three specific reading constructs, including:

- comprehend explicit details (e.g., phrases, sentences) - 3 items
- infer the meaning of a word from context - 2 items
- infer meaning from sentences - 1 item

2.3. Documentation of item writers' intentions

The documentation of item writers' intentions was conducted after the first drafts of the items had been completed. The information charted for each item includes: (1) features of the item, and (2) rationales behind the item's content. Afterwards, the documentation will be grouped according to specific reading constructs.

Regarding item content, the rationales behind the key and the distractors of each item were documented. This information helps clarify how each component of the item contributes to assessing the targeted reading construct.

Regarding item features, the relationship between an item and the text will be coded in terms of text explicitness features, following the taxonomy proposed by Hasegawa (2017). This taxonomy combined two complementary criteria for analysing text explicitness: (1) the size of the relevant text portion, and (2) the type of cognitive activity engaged. The size of the relevant text portion to answer an item can be classified as *within-sentence* (incl. *intra-clause* and *inter-clause*) or *between-sentence* (incl. *adjacent* and *non-adjacent*); The cognitive activity can be categorised as either *textual* (incl. *extracting* and *paraphrasing*) or *inferential* (incl. *bridging* and *elaborating*). Due to the word limits, the full definitions of the concepts in the taxonomy were included in Appendix A.

It should be noted that in Hasegawa's study, the format of the test items was True/False, which only included one option for readers to compare with the text. However, the format of test items in this study is 4-option multiple-choice (MC). Therefore, the analysis of text explicitness is applied to all four options in the item.

2.4. Piloting the testlet

After the testlet had been constructed and the item writer's intentions had been documented, it was piloted on two test-takers. Each of the test-takers did the test in approximately 20 minutes and was invited to join an interview right after their completion of the test. The interview aimed to gain insight into how the test-takers read and selected answers for each item. It is important to note that, given the limited number of test-takers, their data will only be used as

e-ISBN: 978-629-98082-2-0

supplementary comments where relevant in the discussion.

3. Results and discussion

3.1. MC item targeting comprehending explicit details

The documentation shows that MC items designed for comprehending explicit details can be either text-implicit or text-explicit. Specifically, Item 10 is considered text-explicit since the portion of text relevant to each option of the item is within a sentence, and the cognitive process involved in each of them is paraphrasing from explicit textual information. The features of Item 10 are presented in Table 2.

The intention behind the design of Item 10 was to get test-takers read carefully the whole paragraph, and assess their comprehension of explicit information presented within individual sentences. Therefore, each option is a paraphrase of a single sentence in the paragraph. It should be noted that the key (i.e., correct answer) in this item is an incorrect paraphrase of the textual information, and the distractors the correct paraphrases. This design aims to lower the unnecessary cognitive load of the reasoning processes that potentially happen when test-takers need to eliminate many options that conflict with the correct understanding of the text, thereby encouraging them to focus more on comprehension.

Table 2: Item Features of Item 10

Item no.	10	
Stem	According to paragraph VIII, the research suggests all of the following for a more productive debate, EXCEPT____.	
Necessary information		
So how can you apply this idea in your own life (S1)? The research offers a few suggestions (S2). Before a tough conversation, you can think about the most important value to you (S3). This can help you stay open-minded during the talk (S4). If the other person gets emotional, you can try to consider how their behaviour was affected by factors such as stress or tiredness (S5). This will help you to take things less personally (S6). Also remember: what is most important to you might not be as important to them, and that's okay (S7). Understanding this not only leads you to a productive disagreement (S8). It also makes you feel stronger within yourself (S9).		
Key overlooking the emotions of other people	**Size of text**	W-Intra (S5)
	Cognitive processes	T-P
Distractor 1 thinking about your core values before debating	**Size of text**	W-Intra (S3)
	Cognitive processes	T-P
Distractor 2 acknowledging differences in personal values	**Size of text**	W-Intra (S7)
	Cognitive processes	T-P

e-ISBN: 978-629-98082-2-0

Distractor 3	**Size of text**	W-Intra (S5)
considering the reasons behind people's behaviour	**Cognitive processes**	T-P

**Note: W-Intra=Within-sentence Intra-clause, T-P=Textual-Paraphrasing, S=sentence*

The other two items in this group, Item 2 and Item 4, can be deemed as text-implicit items (see Table 3 & Table 4). This is because the keys in both items anticipate the comprehension of two sentences or more; the cognitive processes involved in arriving at the keys include both bridging inferences and paraphrasing.

Item 4 was designed with the intention to get test-takers read a small portion of the text and assess their ability to recognise and comprehend a straightforward piece of information relevant to the question. Distractors 2 and 3 are designed to attract test-takers who might not be able to recognise the relevant information; hence, they were derived from the information occurring in the area of necessary information, but were irrelevant to answering the question. Distractor 1 is derived from the relevant information, yet incorrectly paraphrased; it aimed to attract test-takers who can recognise the relevant information but may misunderstand the message.

Table 3: Item Features of Item 4

Item no.	4
Stem	How did the researchers measure the intellectual humility of each participant, as stated in paragraph IV?
Necessary information	*The researchers later analysed the videos to measure the level of intellectual humility of each participant (S1). To do that, they looked for signs of qualified engagement (S2).*

Key	**Size of text**	B-Adj (S1-S2)
They examine the qualified engagement of the participants	**Cognitive processes**	I-B and T-P
Distractor 1	**Size of text**	B-Adj (S1-S2)
They rate the quality of arguments of the participants	**Cognitive processes**	I-B and T-P
Distractor 2	**Size of text**	W-Intra (S1)
They watched the recordings carefully	**Cognitive processes**	T-P
Distractor 3	**Size of text**	W-Intra (S1)
They tested how smart the participants were	**Cognitive processes**	T-P

**Note: B-Adj= Between-sentence Adjacent, W-Intra=Within-sentence Intra-clause, I-B=Implicit-Bridging, T-P=Textual-Paraphrasing, S=sentence*

e-ISBN: 978-629-98082-2-0

Item 2 was designed to have test-takers read a larger portion of the text, two paragraphs, and assess their understanding of how the key information is connected through explicit cohesive devices. As a result, all the key and distractors include the correct paraphrases of the key information but differ in the connections between the information in terms of temporal sequence.

Table 4: Item Features of Item 2

<table>
<tr><td>Item no.</td><td colspan="2">2</td></tr>
<tr><td>Stem</td><td colspan="2">Which of the following BEST summarises the process that people went through in the study, as described in paragraphs II and III?</td></tr>
<tr><td colspan="3">Necessary information

In the study, 303 students and community members were invited to join a debate about university tuition fees (S1). Before the debate, some people were given a list of 19 values, such as career achievement, caring for others, or protecting nature and asked to write about the one that mattered most to them (S2). The others were, instead, asked to write about a neutral topic (S3).

All participants then watched a presentation that showed arguments for and against tuition fees (S4). After that, they were divided into small groups of two to four people to decide which arguments were the most and the least persuasive (S5).</td></tr>
<tr><td rowspan="2">Key
Step 1: Write about important values OR neutral topics
--> Step 2: Watch arguments
--> Step 3: Group discussion</td><td>Size of text</td><td>B-Adj (S1-S2-S3-S4-S5)</td></tr>
<tr><td>Cognitive processes</td><td>I-B and T-P</td></tr>
<tr><td rowspan="2">Distractor 1
Step 1: Write about important values
Step 2: Write about neutral topics
Step 3: Watch arguments
Step 4: Group discussion</td><td>Size of text</td><td>B-Adj (S1-S2-S3-S4-S5)</td></tr>
<tr><td>Cognitive processes</td><td>I-B and T-P</td></tr>
<tr><td rowspan="2">Distractor 2
Step 1: Write about neutral topics
Step 2: Write about important values
Step 3: Watch arguments
Step 4: Group discussion</td><td>Size of text</td><td>B-Adj (S1-S2-S3-S4-S5)</td></tr>
<tr><td>Cognitive processes</td><td>I-B and T-P</td></tr>
<tr><td>Distractor 3</td><td>Size of text</td><td>B-Adj (S1-S2-S3-S4-S5)</td></tr>
</table>

e-ISBN: 978-629-98082-2-0

Step 1: Watch arguments Step 2: Write about important values OR neutral topics Step 3: Group discussion participants were	**Cognitive processes**	I-B and T-P

**Note: B-Adj= Between-sentence Adjacent, W-Intra=Within-sentence Intra-clause, I-B=Implicit-Bridging, T-P=Textual-Paraphrasing, S=sentence*

3.2. Item targeting inferring the meaning of a word from context

Table 5 below shows the features of two items targeted at the ability to infer the meaning of a word from context.

Table 5: Item features of Item 1 and Item 3

Item no.		1	3
Stem		What is the MOST suitable meaning of 'well-documented' in the context of paragraph I?	What is the MOST suitable meaning of 'persuasive' in the context of paragraph III?
Necessary information		*Past research shows that intellectual humility, being willing to accept that our beliefs might be wrong, can help us have more productive debates. While the benefits of humility have been* ***well-documented****, very few studies have looked at the factors that might increase humility.*	*All participants then watched a presentation that showed arguments for and against tuition fees. After that, they were divided into small groups of two to four people to decide which arguments were the most and the least* ***persuasive****.*
Key-Item 1: supported by much evidence **Key-Item 3:** convincing	**Size of text**	B-Adj	B-Adj
	Cognitive processes	I-B	I-E

**Note: B-Adj= Between-sentence Adjacent, I-B=Implicit-Bridging, I-E=Implicit-Elaborating, S=sentence*

All the items were found to be text-implicit in general. Regarding the size of text relevant to the keys, both items required the comprehension of at least two adjacent sentences. That said, the differences lie in the cognitive processes involved in inferring the meaning of the targeted words from the text, provided that the test-takers do not know the words before taking the test.

To solve Item 1, test-takers need to interpret the clause containing the unfamiliar word 'well-documented', which has a clear referential link to the preceding sentence: the idea expressed in

e-ISBN: 978-629-98082-2-0

the clause refers back to the overall message of that sentence. This prior sentence thus provides a strong contextual clue for inferring the meaning of the unknown word. By understanding both the preceding sentence and its connection to the clause with the target word (i.e., making a bridging inference), test-takers are likely to identify the correct answer.

In Item 3, the clause containing the unfamiliar word has a weaker connection to the preceding sentence. Although the word 'argument' appears in both, the ideas associated with it differ in theme. As a result, the contextual clue is less explicit, and test-takers may need to rely on their background knowledge to make an elaborative inference about the most likely meaning of the word in context (i.e., what people typically decide after hearing arguments?).

The data from test-takers' performance showed that while all of them got the correct answer for Item 1, only one person got the correct answer for Item 3. The test-taker who was attracted by a distractor ('interesing') in Item 3 revealed in the interview that to her, 'interesting' is the most suitable meaning of the word in the context. This somehow supported the rationale that the background knowledge plays a role in inferring the meaning of the targeted word in Item 3, which potentially made Item 3 more difficult than Item 1 despite the similar construct they targeted.

3.3. Item targeting inferring meaning from sentences

Table 6 below shows the features of the item designed to assess the ability to infer the meaning from sentences.

Table 6: Item features of Item 6

Item no.	6	
Stem	As discussed in paragraph V, which of the following can be seen as boosted conviction?	
Necessary information		
	People were also seen as more intellectually humble if they engaged in less boosted conviction (S1). This is linked to a more arrogant style of speaking (S2). It includes the use of words like "obviously" or "always," or statements like "how can you say that?" (S3). These ways of speaking suggest that the speaker thinks their view is clearly right and the other person's view is just wrong (S4).	
Key	**Size of text**	B-Adj (S2-S3-S4)
self-centered statements	**Cognitive processes**	I-B
Distractor 1	**Size of text**	W-Intra (S2)
rude comments	**Cognitive processes**	T-P
Distractor 2	**Size of text**	W-Intra (S3)

e-ISBN: 978-629-98082-2-0

simple words	**Cognitive processes**	I-E
Distractor 3	**Size of text**	W-Intra (S3)
non-sense claims	**Cognitive processes**	I-E

**Note: B-Adj= Between-sentence Adjacent, W-Intra=Within-sentence Intra-clause, I-B=Implicit-Bridging, I-E=Implicit Elaborating, T-P=Textual-Paraphrasing, S=sentence*

Item 6 can be considered as a text-implit question. The key depends on the comprehension of three adjacent sentences and a bridging inference to connect them. However, regarding the distractors, the item writers mostly targeted specific phrases within a sentence. Distractor 1 was derived from the phrase 'arrogant style of speaking', which was incorrectly paraphrased as 'rude comments'. Distractors 2 and 3 were based on elaborative inferences from certain phrases that might make sense if the phrases are taken out of the paragraph's context, like 'simple words' can be inferred from 'obviously' and 'always'.

One test-taker was attracted by Distractor 3 ('simple words'). Later, she shared in the interview that she read fast to answer this question. Fast reading, hence, is likely to affect her complete comprehension of the necessary information, leading to an elaborative inference from a small part of the necessary information instead of inferring from the whole of it.

4. Conclusion

The study presented a detailed documentation of item writers' intentions when writing MC items to assess different reading constructs. Regarding item features, despite the different constructs the items targeted at, five out of six items were identified as text-implicit, which requires the comprehension of two or more sentences and making bridging inferences to get the correct answers. Besides, the size of texts and cognitive activities involved in each option within the same item can be different from each other. The documentation of those features, together with rationales behind the item content, seems to hold a promise for providing explanations as to why the test-takers might be attracted to certain distractors, replacing a general assumption of wrong comprehension.

Despite its detailed documentation of item writers' intentions, the study has certain limitations. In terms of participants, the number of test-takers was very small, which limits the strength of any claims drawn from their data and makes it difficult to rigorously compare their responses with the item writers' intentions. As for data analysis, the coding of item features was carried out by the original item writer of a specific item without a second rater. However, this was unavoidable since the original item writer is the one who designed and determined how the items should work, thereby being the most accurate coder of their items.

e-ISBN: 978-629-98082-2-0

References

Alderson, J. C. (1990). Testing reading comprehension skills (Part two). Getting students to talk about taking a reading test. (A pilot study). *Reading in a Foreign Language*, *7*(1), 465-503.

Green, A., & Hawkey, R. (2011). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing*, *29*(1), 109-129. https://doi.org/10.1177/0265532211413445

Green, A., & Hawkey, R. (2012). An empirical investigation of the process of writing Academic reading test items for the International English Language Testing System. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 270-378). Cambridge: Cambridge University Press.

Hasegawa, Y. (2017). Analyzing Explicit and Implicit Reading Questions in a Term-Exam: A Case Study. *JLTA Journal*, *20*, 57-75.

Jeon, E. H., & Yamashita, J. (2020). Measuring L2 reading. In P. Winke & T. Brunfaut (Ed.), *The Routledge handbook of second language acquisition and language testing* (pp. 265-274). New York: Routledge.

Jones, G. (2020). Designing multiple-choice test items. In P. Winke & T. Brunfaut (Ed.), *The Routledge handbook of second language acquisition and language testing* (pp. 90-101). New York: Routledge.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullion, V. (2010). A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual

Liu, X. (2018). *Establishing the Foundation for a Diagnostic Assessment of Reading in English for Academic Purposes*. [Unpublished doctoral thesis], University of Auckland.

Liu, X., & Read, J. (2021). Investigating the Skills Involved in Reading Test Tasks through Expert Judgement and Verbal Protocol Analysis: Convergence and Divergence between the Two Methods. *Language Assessment Quarterly*, *18*(4), 357–381. https://doi-org.ezproxy.lancs.ac.uk/10.1080/15434303.2021.1881964

Ngo, P. V. H. (2020). *Understanding Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests: A Case of English as a Foreign Language (EFL) Teachers in Vietnamese Higher Education Settings*. [Unpublished doctoral thesis]. University of Melbourne.

Rossi, O. (2021). *Item writing skills and their development: Insights from an induction item writer training course.* [Unpublished doctoral thesis]. Lancaster University.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441-474. https://doi.org/10.1191/0265532206lt337oa.

Salisbury, K. (2005). *The Edge of Expertise: Towards an Understanding of Listening Test Item*

e-ISBN: 978-629-98082-2-0

Writing as Professional Practice. [Unpublished doctoral thesis]. King's College London, UK.

Schedl, M., & Malloy, J. (2013). Writing items and tasks. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 796-813). London: John Wiley & Sons.

Appendix A

Hasegawa's taxonomy of item classification based on text explicitness (Hasegawa, 2017, p.66)

Table 1

Descriptions of Two Textual and Two Inferential Types of Questions

Item type	Subtype	Description
Textual (Explicit)	Extracting	Questions that can be solved by extracting a specific piece of information explicitly written in the text.
	Paraphrasing	Questions that can be solved by identifying a specific piece of information explicitly written in the text and paraphrasing it.
Inferential (Implicit)	Bridging	Questions that can be solved by bridging inferences such as understanding referential and causal relationships among pieces of information.
	Elaborating	Questions that can be solved by elaborative inferences to interpret what is not explicitly written in the text.

Table 2

Descriptions of Two Within-Sentence and Two Between-Sentence Types of Questions

Item type	Subtype	Description
Within-sentence (Explicit)	Intra-clause	Questions that can be solved by comprehending a specific clause in a specific sentence.
	Inter-clause	Questions that can be solved by comprehending at least two clauses in a specific sentence.
Between-sentence (Implicit)	Adjacent	Questions that cannot be solved without comprehending two adjacent sentences and integrating the information.
	Nonadjacent	Questions that cannot be solved without comprehending at least two nonadjacent sentences and integrating the information.

e-ISBN: 978-629-98082-2-0

P053

TOEFL Multi-Faceted Validity in an Indonesian Context: A Systematic Review

Dian Purwitasari

University of Limerick, Ireland
dianpurwita2@gmail.com

Abstract

As a standardised test, it is undoubtedly true that TOEFL guarantees quality and validity in measuring English proficiency. However, Messick (1989) sees validity as a more complex concept, not only considering the construct validity, but also the utility and relevance, value implications, and social consequences of testing. Considering this concept, this research investigates what scholars have discovered in regards to TOEFL practices in Indonesia to reveal its validity evidence using a systematic review approach. This study identifies the multi-faceted validity of TOEFL in Indonesia as shown by various evidence. This ranged from how this test is taught in classes, and how it is done by the test-takers and used by the gatekeepers. Generally, TOEFL tests have resulted in numerous intense difficulties, which are revealed in 17 studies, and have led to test-takers' anxiety in sitting the test (Muliawati et al, 2020; Sunarti et al, 2019). To help the test-takers, institutions have provided TOEFL preparation classes applying various teaching methods; this is discussed in eight studies. Meanwhile, seven studies explored the policy implementing TOEFL in institutions and what the students' perceptions are of the policy. In conclusion, it is evident that TOEFL practices possess multi-faceted validity reflected from various evidence. TOEFL on its own has well-structured components which assess test-takers' English ability and this test also affects the users in different social spheres. However, TOEFL practices in Indonesia are not without constraint and further supports for test-takers in advancing their English proficiency are required.

Keywords: Validity; validity evidence; TOEFL

1. Introduction

The foundation of language testing relies on what is referred to as 'validity' which indicates the quality of the test. Validity is associated with the understanding of what a test measures is what it is supposed to test. It also describes the degree to which theory and evidence underpin the test score interpretation for its use (Chapelle & Lee, 2022). To a larger extent, Bachman & Palmer (1996) argued that in considering the specific qualities that determine the overall usefulness of a test, it is worth noting that tests are parts of a larger societal or educational

e-ISBN: 978-629-98082-2-0

context.

TOEFL has been used worldwide and its validity has been extensively studied in different contexts, for example Chapelle (2008), Harsch et al. (2017) and O'Dwyer et al. (2018). Likewise, Indonesian scholars have studied TOEFL from various points of view and settings. This study seeks the validity evidence of TOEFL practices in Indonesia to examine the pattern of its societal, political, and educational contexts. The current research addresses the following research question 'what are the validity evidence of TOEFL in Indonesia?'. This research aims to draw a general understanding of TOEFL practices in the country based on the evidence of validity by Messick (1989).

2. Methods

To achieve the objective of the research, the method applied to this study is a systematic review which involves the study of the existing research on the given topic. Pollock & Berge (2017) defined a systematic review as the identification of the primary research relevant to the given review question, the critical appraisal, and the synthesis of finding. Aromataris & Pearson (2014) mentioned that systematic review is not intended to create a new theory or concept but to synthesize and summarise the existing knowledge.

This study set the inclusion criteria for the journal articles published between the end of 2014 to early 2025, while the excluded studies are those published in the Indonesian language. The data selection began with the keywords 'TOEFL Validity Indonesia' typed into the Google Scholar search engine, which was redirected to Indonesian open access journals. The critical appraisal involved checking the journal articles which were all indexed and peer-reviewed. The next step was to screen the abstract for the data extraction, followed by summarising and categorising the papers with similar topics to reveal their patterns indicating TOEFL validity evidence.

3. Result and discussion

The data collection resulted in 41 studies included in the review. The main themes are as follows:

Table 1: Finding

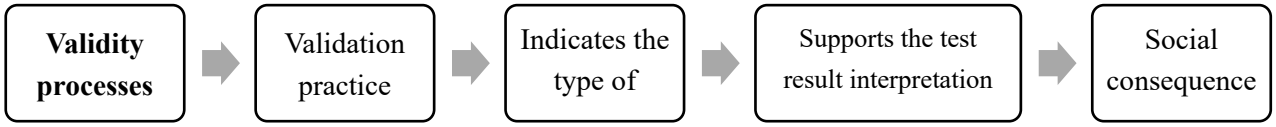
Main themes	Studies by	Qt	%
Language policy and students' perception	(Kaniadewi, 2023; Karjo & Ronaldo, 2018; Khoiruman & Irawan, 2025; Munandar, 2019; Raharjo, 2020; Rahma et al., 2021; Zakiah et al., 2023)	**7**	**17.07%**
Content analysis	(Dewi et al., 2023; Ismail et al., 2019; Taufiq et al., 2018; Tika et al., 2023)	**4**	**9.75%**

e-ISBN: 978-629-98082-2-0

Test-taker difficulty	(Akmal et al., 2020; Bania, 2024; Fitri, 2017; Fitria, 2021; Gunantar & Rosaria, 2023; Halim & Ardiningtyas, 2018; Hampp et al., 2021; Jasrial et al., 2022; Karimullah & Mukminatien, 2022; Mahmud, 2014; Maulana & Lubis, 2022; Muliawati et al., 2020; Purba et al., 2024; Setiawati et al., 2024; Sunarti et al., 2019; Yoestara & Putri, 2019; Yosintha et al., 2021)	**17**	**41.46%**
Correlation with different skills and achievement	(Asmani, 2014; Hutabarat, 2023; Karjo & Andreani, 2018; Liskinasih & Lutviana, 2016; Setyowati et al., 2020)	**5**	**12.19%**
Classroom application	(Dalimunte et al., 2025; Kaniadewi & Asyifa, 2022; M. S. Maharani & Putro, 2021; S. Maharani & Miftachudin, 2021; Maryansyah et al., 2023; Parmadi & Kepirianto, 2023; Pratiwi et al., 2021; Syakur et al., 2019)	**8**	**19.51%**

The inquiry of language test validity is not as simple as finding the effectiveness of a testing instrument in measuring a particular ability. Messick (1989) proposed a complex conception of validity and validation processes which provided robust recommendation for validation practice by indicating the types of evidence that can be used by researchers to support the interpretations and the uses of test score. To summarise, Young (2022) divided two dimensions of language testing, namely: (1) the construct of language knowledge in which the test is designed and which result is interpreted, and (2) the social consequences of testing.

Figure 1: the steps of validity processes using (Messick, 1989) and (Young, 2022)



In table 2, these points are summarised along with the connection to the research themes found in this study to disclose the test's score interpretation for a specific use. This concept has been extensively discussed in validation research as it includes both the test construct and its value implication, and social consequences.

Table 2: The types of validity evidence (Messick, 1989) and themes found in the review

Types of evidence (Messick, 1989)	**Related themes found in the review**
Rationales and experts' judgement about the content of the test	Content analysis
The research on the test-takers' response processes	Test-taker difficulty
Quantitative exploration in the internal structure of the response data	Content analysis
Correlations to other variables	Correlation with different skills and achievements
The consequences of testing	Language policy and students' perception, classroom application

e-ISBN: 978-629-98082-2-0

3.1 Rationales and expert judgement about the content of the test

This evidence revealed the match or mismatch between the test content intended by the test developers and evaluation from the expert (Messick, 1989). Ismail et al. (2019) explored TOEFL teaching materials in Barron's, Cambridge, Cliff's, and Longman in examining cultural biases which are found in the Reading Section. Despite the small quantity of cultural biases found, this was not alarming and TOEFL still objectively measured academic English as intended.

Taufiq et al. (2018) and Tika et al. (2023) concluded that TOEFL ITP was highly practical due to its low cost and test structure with multiple choice questions. Tika et al. (2023) added that this test is highly practical for institutions as it allows them to run their own tests. However, it is advised for institutions to conduct a separate test for the test-takers' productive skills.

3.2 The research on the test-takers' response processes

This evidence focuses on the processes involving test-takers ability intended by the test task design and the stated construct interpretation (Messick, 1989).

3.2.1 Difficulty in obtaining minimum scores

Bania (2024) discovered that lecturers and students obtained less than 500 in TOEFL Prediction. This score is below than what is required for doctoral studies and graduation requirements resulting in participants to retake the test.

3.2.2 Difficulty in Structure and Written Expression

Akmal et al. (2020), Fitri (2017), Hampp et al. (2021) and Yosintha et al. (2021) discovered the difficulties faced by test-takers specifically in Structure and Written Expression. Hampp et al. (2021) emphasised that students were mostly unable to answer error analysis. Yosintha et al. (2021) added that the most difficult questions involve active-passive verbs, double comparatives, and pronoun-noun agreements.

3.2.3 Difficulty in Listening Sections

Fitria (2021) discovered that students found the Listening section challenging due to the speakers' accent and speed, and the students tend to be lacking practice and vocabulary.

3.2.4 Difficulties found in all parts of the test

Gunantar & Rosaria, (2023); Jasrial et al., (2022); Maulana & Lubis, (2022); and Setiawati et al. (2024) agreed that Structure and Written Expression is the most challenging of all three sections, followed by Listening and Reading.

e-ISBN: 978-629-98082-2-0

3.2.5 Factors affecting the difficulties

Halim & Ardiningtyas (2018), Mahmud (2014) and Purba et al. (2024) revealed that there were numerous factors affecting the difficulties faced by Indonesian test-takers. These included low English proficiency, insufficient practices, low motivation, and lack of support from tutors.

3.2.6 Difficulties in iBT TOEFL

Karimullah & Mukminatien (2022) explored TOEFL iBT IRLW and discovered that students encountered various difficulties, including grammar and paraphrasing. The test-takers also applied different integrated writing strategies incorporating their learning style, and cognitive levels.

3.2.7 Anxiety and students' self-efficacy

Muliawati et al. (2020) claimed that 80% of students experienced moderate anxiety while the remaining experienced mild anxiety before and during the test. Sunarti et al. (2019) added that the lower the anxiety faced by the test-takers, the higher the TOEFL score they obtained. Yoestara & Putri (2019) found that 77% of students possessed medium level of self-efficacy but only 25% could achieve 450 on their TOEFL score. It indicated low positive correlation between students' self-efficacy and TOEFL performances.

3.3 Quantitative exploration in the internal structure of the response data

This statistical evidence is expected to show whether or not the responses provided by the test-takers are consistent with the intended interpretation of the dimensionality of the test construct (Messick, 1989). Ockey (2022) stressed that relating test-taker ability to the items' level of difficulty makes it possible to mathematically model the probability that a test-taker is able to respond appropriately to a test item.

Using the Rasch Model in analysing the Reading section, Dewi et al. (2023) found that 36.8% of the items were found to be too easy or too difficult to be given to the target test-takers, while the rest were considered to have followed the standard.

3.4 Correlations to other variables

This evidence connected the relationships expected between the construct measured by the test and the measures of other variables (Messick, 1989). A study conducted by Setyowati et al. (2020) found a correlation between students' vocabulary level and reading comprehensions in TOEFL. In general, the better vocabulary mastery possessed by the test-takers, the higher the score they achieved. Similarly, Asmani (2014) discovered that students' Listening scores correlated significantly with Business English speaking skills with a moderate linear relationship in TOEFL iBT.

e-ISBN: 978-629-98082-2-0

In addition to language proficiency, TOEFL also indicates test-takers' achievement in educational settings. Karjo & Andreani (2018) discovered that the students' entry score of TOEFL can predict their exit score. On the other hand, TOEFL cannot predict students' academic achievement but has a positive correlation between the two. Similarly, Liskinasih & Lutviana (2016) added that there was a strong positive correlation between TOEFL score as the placement test and students' achievement in the Integrated English Course. The higher the TOEFL score, the higher the students' score in the course.

3.5 The consequences of testing

The discussion on the consequences of testing is derived from test score interpretations which carry social implications to the test-taker. Messick (1989) oversaw whether the intended impact of testing results for the test-takers, users, and other related parties were indirectly affected by the application of the test. Young (2022) emphasises this consequence as the social dimension of language testing.

3.5.1 Social consequence

Munandar (2019) reported that TOEFL influences institutions to include TOEFL preparation which students favour more compared to regular English classes. TOEFL also affected individuals in providing better employment opportunities. This claim was supported by Khoiruman & Irawan (2025) who were convinced that TOEFL scores showcased potential for better communication in English which attracts employers, especially multi-national companies.

3.5.2 Students' perspective

Kaniadewi (2023), Rahma et al. (2021) and Zakiah et al. (2023) reported that students generally welcomed the policy using TOEFL as a graduation requirement. However, they mentioned that many students were unsuccessful in obtaining the minimum score. Raharjo (2020) added that the students asked for other types of assessment, such as portfolios, to be considered.

3.5.3 Classroom application

Kaniadewi & Asyifa (2022) discovered that TOEFL Preparation programmes have improved the students' score significantly. Dalimunte et al. (2025) ascertained that various methods of teaching helped to improve the students' tests. In addition, S. Maharani & Miftachudin (2021) mentioned that there were 10 methods used by teachers in teaching TOEFL classes, with the most commonly used being drilling and practice, and discussion.

In addition to conventional methods of teaching, Pratiwi et al., (2021) reported that different E-learning approaches including Quizizz and Kahoot improved students' learning outcomes in Structure and Written Expression. This finding is strengthened by Syakur et al. (2019) stating

e-ISBN: 978-629-98082-2-0

that the application of e-learning methods using LMS is proven to have improved EFL learners' TOEFL score overall.

4. Conclusion

The exploration of validity evidence in Indonesia has revealed that most Indonesian test-takers seemed to have faced higher degrees of difficulty which are reflected in 17 studies (41.46%). Furthermore, the study conducted by Dewi et al. (2023) revealed that TOEFL Reading items at a moderate difficulty may strengthen the assumption that Indonesians generally have low levels of English proficiency. This evidence is followed up by the efforts made by related institutions, where tutors applied various teaching methods to help students obtain the TOEFL score they require. Eventually, the presence of TOEFL in Indonesian society allows for the interaction between test-takers, users, schools, and stakeholders resulting in various degrees of social consequences which strengthen the power of language testing as quoted by Young (2022).

Acknowledgement

I'm very grateful for the support given to me by my former thesis supervisor Dr Elaine Riordan who introduced me to systematic review. Thank you to my partner Jim Purcell who has been very patient in assisting me during my hardship with a health condition, as some part of this paper was written in a hospital bed.

References

Akmal, S., Rasyid, M. N. A., Masna, Y., & Soraya, C. N. (2020). EFL learners' difficulties in the structure and written expression section of TOEFL test in an Indonesian university. *Englisia: Journal of Language, Education and Humanities*, *Vol. 7, No. 2*, 156–180. https://doi.org/10.22373/ej.v7i2.6472

Aromataris, E., & Pearson, A. (2014). Systematic reviews, Step by Step. *The Joanna Briggs Institute*, *114*(3), 53–58.

Asmani, A. B. (2014). Correlative analysis of TOEFL iBT scores of listening skill versus scores of Business English speaking skill among Binus University sophomores. *Jurnal Lingua Cultura*, *Vol.8 No.2*, 85–94.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford University Press.

Bania, A. S. (2024). Evaluating TOEFL Prediction test proficiency among lecturers and students at the University of Samudra. *English Review: Journal of English Education*, *12*(1), 157–166. https://doi.org/10.25134/erjee.v12i1.8594

Chapelle, C. A. (2008). The TOEFL Validity Argument. In *Building a Validity Argument for the*

e-ISBN: 978-629-98082-2-0

Test of English as a Foreign Language (1st ed.). Routledge.

Chapelle, C. A., & Lee, H. (2022). Conceptions of validity. In *The Routledge Handbook of Language Testing* (2nd ed., pp. 17–31). Routledge.

Dalimunte, A. A., Somanawattana, C., & Siregar, Y. (2025). The effectiveness of TOEFL program on the students' English proficiency performance: A case of an Indonesian university. *Journal of English Education and Linguistics Studies*, *12(1)*, 144–177. https://doi.org/10.30762/jeels.v12i1.3970

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education)*, *9(1)*, 24–36. https://doi.org/10.21831/reid.v9i1.53514

Fitri, M. (2017). The difficulties faced by the students in answering the Written Expression section of the TOEFL test. *Indonesian Journal of Integrated English Language Teaching*, *3*(2), 206–215.

Fitria, T. N. (2021). An Analysis of the Students' Difficulties in TOEFL Prediction Test of Listening Section. *ENGLISHFRANCA: Academic Journal of English Language and Education*, *Vol. 5, No.1*, 95–110. http://dx.doi.org/10.29240/ef.v5i1.2212

Gunantar, D. A., & Rosaria, S. D. (2023). Difficulties of non-English study program students in carrying out the Institutional TOEFL Test. *English Education: Jurnal Tadris Bahasa Inggris*, *16(1)*, 119–134.

Halim, N., & Ardiningtyas, S. Y. (2018). Difficulties faced by the students in answering TOEFL test questions. *English Teaching Learning and Research Journal*, *Vol.4 No.2*, 219–231. https://doi.org/10.24252/Eternal.V42.2018.A7

Hampp, P. L., Lengkoan, F., & Kumayas, T. A. (2021). Synthesizing grammar and structure problems faced by Indonesian TOEFL participants. *Jurnal Pendidikan Bahasa Inggris Undiksha*, *Vol. 9, No. 1*, 64–68. http://dx.doi.org/10.23887/jpbi.v9 i1.33811

Harsch, C., Ushioda, E., & Ladroue, C. (2017). Investigating the predictive validity of TOEFL iBT® test scores and their use in informing policy in a United Kingdom university setting. *ETS Research Report Series*, *17*(1), 1–80. https://doi.org/10.1002/ets2.12167

Hutabarat, P. (2023). Comparing TOEFL and teacher's assessment scores as a snapshot of student's English skills. *Journal of English Education and Teaching (JEET)*, *Volume 7 number 4*, 770–790.

Ismail, N. M., Mukti, A., Yoestara, M., Delavari, H., & Qadiri, A. (2019). Revisiting cultural biases in TOEFL test: A pilot study. *Proceedings of the 2nd English Education International Conference (EEIC)*, 96–101.

Jasrial, D., Yunita, W., & Villia, A. S. (2022). Exploring the Indonesian nursing students' sifficulties in answering the TOEFL Prediction test. *Metathesis: Journal of English Language Literature and Teaching*, *Vol. 6, No. 2*, 213–224. https://doi.org/10.31002/metathesis.v6i2.97

e-ISBN: 978-629-98082-2-0

Kaniadewi, N. (2023). The analysis of mandatory TOEFL Test policy in University of Muhammadiyah Prof. Dr. HAMKA. *Journal of English Teaching*, *Vol. 9, No. 3*, 323–334. https://doi.org/10.33541/jet.v9i3.4981

Kaniadewi, N., & Asyifa, D. I. (2022). The effect of TOEFL Preparation Course on EFL undergraduate students' on TOEFL scores. *Linguistic, English Education and Art (LEEA) Journal*, *Vol. 6, No. 1*, 12–20. https://doi.org/10.31539/leea.v6i1.4365

Karimullah, I. W., & Mukminatien, N. (2022). Problems faced and strategies applied by test-takers in completing the TOEFL iBT test. *Studies in English Language and Education*, *9(2)*, 574–590. https://doi.org/10.24815/siele.v9i2.23129

Karjo, C. H., & Andreani, W. (2018). Is TOEFL compulsory for Indonesian university students? A correlational study between TOEFL scores and students' academic achievement. *Konferensi Linguistik Tahunan Atma Jaya 16*, 111–115.

Karjo, C. H., & Ronaldo, D. (2018). The validity of TOEFL as entry and exit college requirements: *Advances in Social Science, Education and Humanities Research*, *254*(Eleventh Conference on Applied Linguistics (CONAPLIN 2018)), 326–330.

Khoiruman, M. A., & Irawan, D. H. (2025). Analysing the role of English Language Proficiency (TOEFL) in increasing job opportunities in the global industry sector. *Huele: Journal of Applied Linguistics, Literature and Culture*, *Vol. 5, No. 1*, 1–15.

Liskinasih, A., & Lutviana, R. (2016). The validity evidence of TOEFL test as placement test. *Jurnal Ilmiah Bahasa Dan Sastra*, *Vol. 3, No. 2*, 173–180.

Maharani, M. S., & Putro, N. H. P. S. (2021). Evaluation of TOEFL preparation course program to improve students' test score. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *Vol 25, No 1*, 63–76. https://doi.org/10.21831/pep.v25i1.39375

Maharani, S. & Miftachudin. (2021). Revealing teachers' methods in teaching Test of English as a Foreign Language (TOEFL). *Journal of English Teaching and Learning Issues*, *4(2)*, 119–130. https://doi.org/10.21043/jetli.v4i2.12204

Mahmud, M. (2014). The EFL students' problems in answering the Test of English as a Foreign Language (TOEFL): A atudy in Indonesian context. *Theory and Practice in Language Studies*, *Vol. 4, No. 12*, 2581–2587. https://doi.org/10.4304/tpls.4.12.2581-2587

Maryansyah, Y., Hamzah, S., & Hadiwinarto. (2023). An evaluation on TOEFL workshop program using Stake's Countenance Model. *Exposure 32: Jurnal Pendidikan Bahasa Inggris*, *Volume 12 (1)*, 32–46.

Maulana, C., & Lubis, I. A. (2022). Students' problem of taking TOEFL test at STMIK Royal academic year 2020/2021. *Journal of Science and Social Research*, *1*, 5–10.

Messick, S. (1989). Validity. In *Educational Measurement* (3rd ed., pp. 13–103). Macmillan Publishing.

Muliawati, I., Ismail, N. M., Rizka, B., & Lismalinda. (2020). Test-taking anxiety among EFL university students in TOEFL test: A case study from Indonesian context. *Humanities &*

e-ISBN: 978-629-98082-2-0

Social Sciences Reviews, *8*(3), 200–208. https://doi.org/10.18510/hssr.2020.8321

Munandar, I. (2019). Critical research on the pedagogical, individual, and social impact of the TOEFL PBT introduction as a testing instrument. *Englisia*, *Vol. 6, No. 2*, 117–129. http://dx.doi.org/10.22373/ej.v6i2.4547

Ockey, G. J. (2022). Item response theory and many-facet Rasch measurement. In *The Routledge Handbook of Language Testing* (2nd ed., pp. 462–476). Routledge.

O'Dwyer, J., Kantarcioglu, E., & Thomas, C. (2018). An Investigation of the Predictive Validity of the TOEFL iBT® Test at an English-Medium University in Turkey. *ETS Research Report Series*, *18*(1), 1–13. https://doi.org/10.1002/ets2.12230

Parmadi, A. R., & Kepirianto, C. (2023). Indonesian EFL's learning strategies and personality types in achieving TOEFL score above 500. *Jurnal Pendidikan Bahasa Inggris Undiksha*, *Vol. 11, No. 1*, 18–23. https://doi.org/10.23887/jpbi.v11 i1.57162

Pollock, A., & Berge, E. (2017). How to do a systematic review. *International Journal of Stroke*, *Vol. 13(2)*, 138–156. https://doi.org/10.1177/1747493017743796

Pratiwi, D. I., Atmaja, D. S., & Prasetya, H. W. (2021). Multiple e-learning technologies on practicing TOEFL structure and written expression. *Journal of English Educators Society*, *6*(1), 105–115. https://doi.org/doi: 10.21070/jees.v6i1.1194

Purba, D., Panjaitan, J., & Pardede, D. L. (2024). Problems found on the TOEFL test questions by the fifth semester physics students of Darma Agung University. *Jurnal Darma Agung*, *Vol. 32, No. 2*, 576–582. https://dx.doi.org.10.46930/ojsuda.v32i2.4321

Raharjo, S. D. (2020). Students' perception: Assessing English competence in TOEFL as a standardized English language proficiency test in Indonesian's higher education. *Intensive Journal*, *3(2)*, 40–48.

Rahma, E. A., Syafitri, R., Syahputri, V. N., & Parlindungan, F. (2021). An Evaluation of TOEFL Benchmark Policy as an Exit Requirement for Undergraduate Students. *Southeast Asia Language Teaching and Learning Journal*, *Volume 4 Number 1*, 18–25. https://doi.org/10.35307/saltel.v4i1.61

Setiawati, S. A. P., Yashinta, F., & Chyndy, F. (2024). Difficulties of non-English study program students in taking the TOEFL-Like test at Universitas Muhammadiyah Yogyakarta: A case study of UMY Economics Study Program students batch 2022. *SHS Web of Conferences*. https://doi.org/10.1051/shsconf/202420204002

Setyowati, M., Latifa, A. K., Pratiwi, E., & Mabagits, S. (2020). Student's vocabulary mastery on TOEFL test: Does it correlate with reading comprehension? *Advances in Social Science, Education and Humanities Research*, *434*(International Conference on English Language Teaching (ICONELT 2019)), 245–259.

Sunarti, S., Khatimah, K., & Rachman, D. (2019). The relationship between TOEFL anxiety and TOEFL performance among EFL learners. *Education Journal*, *8(6)*, 296–300. https://doi.org/10.11648/j.edu.20190806.19

e-ISBN: 978-629-98082-2-0

Syakur, A., Junining, E., & Sabat, Y. (2019). Application of E-Learning as a method in educational model to increase the TOEFL score in higher education. *Journal Of Development Research*, *3(2)*, 111–116. https://doi.org/10.28926/jdr.v3i2.88

Taufiq, W., Santoso, D. R., & Fediyanto, N. (2018). Critical analysis on TOEFL ITP as a language assessment. *Advances in Social Science, Education and Humanities Research*, *125*, 226–229.

Tika, A., Dalimunte, A. A., Dalimunte, M., Tanjung, A. P., & Suryani, I. (2023). Evaluating TOEFL ITP test: A critical review. *Humanitatis :Journal of Language and Literature*, *Vol.9 No.2*, 245–254. https://doi.org/10.30812/humanitatis.v9i2.2329

Yoestara, M., & Putri, Z. (2019). University students' self-efficacy: A contributing factor in TOEFL performance. *Studies in English Language and Education*, *6(1)*, 117–130. https://doi.org/10.24815/siele.v6i1.12132

Yosintha, R., Yunianti, S. S., & Ramadhika, B. (2021). Structure and written expressions of the TOEFL: linguistic and non-linguistic constraints. *NOBEL: Journal of Literature and Language Teaching*, *12*(1), 70–90. https://doi.org/10.15642/NOBEL.2021.12.1.70-90

Young, R. F. (2022). Social dimensions of language testing. In *The Routledge Handbook of Language Testing* (pp. 63–80). Routledge.

Zakiah, A., Dahnilsyah, & Masyhur. (2023). Students' perception toward TOEFL prediction test as one of the graduation requirements from university. *International Journal of Educational Best Practices (IJEBP)*, *7*(2), 273–287. https://doi.org/10.32851/ijebp.v7n2.p273-286

e-ISBN: 978-629-98082-2-0

P061

Investigating the Washback Effect of Introductory Arabic Language Writing Assessment on Learning

***Ainul Rasyiqah binti Sazali[1], Norhaslinda binti Hassan[2]**

[1,] Akademi Pengajian Bahasa, UiTM Shah Alam, MALAYSIA,
[2] Akademi Pengajian Bahasa, UiTM Permatang Pauh, MALAYSIA.
(E-mail: [1]ainul437@uitm.edu.my, [2]haslinda.hassan@uitm.edu.my)
**corresponding author: [1]ainul437@uitm.edu.my*

Abstract

This preliminary study aims to explore the washback effect of an Introductory Arabic Language Writing Assessment (IALWA) on student learning outcomes. The concept of washback refers to the influence of testing on teaching and learning, often shaping students' behaviours, motivations, and study practices. This study examines the impact of IALWA on students' approach to language acquisition, with a particular focus on writing skills. By analysing students' perceptions of IALWA, the study evaluates how the assessment aligns with learning outcomes and its role in reinforcing language development. Through qualitative data collection method, which is semi-structured individual interviews, this research seeks to uncover both positive and negative washback effects together with the mediating factors that promote and inhibit the washback. The study identified both positive and negative washback effects. Positive effects were associated with clear teaching practices, formative tools like dummy tests, and student motivation while negative effects stemmed from technical limitations and students' lack of prior knowledge. The findings aim to inform educators and curriculum developers about the potential implications of assessment design, offering insights into how assessments can be structured to promote deeper and more comprehensive language learning experiences. Educators and curriculum designers should ensure that assessment tasks are clearly communicated, aligned with instructional content, and supported by robust preparatory activities in order to maximize positive washback. Moreover, logistical challenges—particularly in digital settings—should be minimized to prevent external factors from undermining assessment reliability. This study concludes by offering recommendations for integrating assessments that promote long-term language acquisition, ensuring alignment between assessment practices, educational goals, and student learning needs.

Keywords: washback; arabic language; writing assessment; language acquisition

e-ISBN: 978-629-98082-2-0

1. Introduction

Good tests acquire three qualities; validity, reliability and practicality. Validity refers to whether a test measures what it intends to measure. The perception of reliability is defined as 'the consistency of measurement' (Bachman, L.F., A.S Palmer, 1996). Reliability should focus on students' score that the students should produce the same scores regardless of how many versions of test the candidates take. The focal point of practicality is management system. It is closely related to cost-cutting measures in time and in money (Hossain, 2015).

Other than these three qualities, there are remaining two principles of language assessment which are authenticity and washback. Authenticity deals with how well the characteristics of the test correlate to the native language (target language); how likely the language tasks will actually be performed in the real world. The last, washback is the feedback that a test gives to both the test takers and the test developers (Brown, 2004). Among the five principles of language assessment, washback contributes more on teaching and learning. In the area of language assessment, the influence of tests or examinations on teaching and learning is known as washback (Alderson & Wall, 1993). Gipps (1994) includes the washback on teaching and the curriculum under the consequential validity. Washback and backwash carry the same meaning. However, backwash is popularly used in general education context while washback is widely used among language testing scholars.

Washback can motivate both students and foreign language instructors to fulfill their teaching and learning goals (Alderson & Wall, 1993). Furthermore, washback may help improving foreign language instructors' creativity in designing and utilizing good tests as beneficial teaching-learning activities in order to encourage a positive teaching-learning process (Pearson, 1988). A creative and innovative test can quite advantageously result in a syllabus alteration or a new syllabus. Other impacts of washback deal with the selection of teaching methods and students' learning styles (Tayeb, et al., 2014). Hence, washback influences the components of the language teaching-learning processes and affects what and how the teachers teach, and the learners learn.

Writing can be conceptualized as a linguistic, cognitive, social and cultural phenomenon. Writing assessment to monitor students' progress is always implemented in foreign language classroom. For washback effect to be effective in improving the student's writing ability, the test of writing should be designed according to the identification of the ability we are intending to test. This in turn requires identifying the factors other than the ability we are intending to test that may be engaged by the test task, so that we can attempt to control them to ensure that the

e-ISBN: 978-629-98082-2-0

inferences about language ability we make on the basis of test results are valid. However, the degree to which a writing test is specifically measuring language as opposed to measuring other cognitive skills is not always clear cut (Kolahi, 2007).

According to Brown (2004), there are four types of writing performance that can be assessed by the writing teachers and instructors; imitative, intensive, responsive and extensive writing. Imitative writing is one of the writing performances that requires students to be able to produce simple written language, such as writing letters, words, punctuation, and very brief sentences. While intensive (controlled) writing assesses the students' skills in producing appropriate vocabulary within a context, collocations, and idioms. To assess these skills, the writing teachers and instructors may employ such tasks of dictation and dicto-comp, grammatical transformation tasks, picture-cued tasks, vocabulary assessment tasks, and short-answer and sentence completion tasks.

Finally, the last two writing performance is responsive and extensive writing. They are regarded as a continuum of possibilities ranging from lower-end tasks which is more complex the previous category (imitative and intensive), through more open-ended tasks such as writing short reports, essays, summaries, and responses, up to texts of several pages or more. Paraphrasing, guided question and answer, and paragraph construction tasks are the assessments for these performance (Nopita, 2019). The first (imitative) and the second (intensive) are the type of writing performance employed in the recent study; Introductory Arabic Writing Assessment.

Bachman and Palmer (1996), discuss two main purposes for language tests. The primary purpose is to make inferences about language ability, and the secondary purpose is to make decisions based on those inferences. These inferences are then used as data for making a variety of decisions at an individual, classroom, or program level. It is possible to make three types of inferences on the basis of a language test: proficiency, diagnosis, and achievement. It is noteworthy that the primary purpose of a language test is to make inferences about language ability. Thus, the focus of the present study is the writing ability, one of the most fundamental concerns in developing a writing test. Due to the complex nature of writing as social, cultural and cognitive phenomenon, and the assortment of challenges faced by both learners and teachers, this study aims to investigate the washback effects of an Introductory Arabic Language Writing Assessment (IALWA) on learning.

e-ISBN: 978-629-98082-2-0

2. Methodology

This qualitative study is done to investigate how Introductory Arabic Language Writing Assessment (IALWA) influence the students' learning. The purpose of this study is to examine the students' perception of IALWA and their learning processes. On top of that, this study also seeks to observe the mediating factors that either promote or inhibits the washback effect. The present study will employ the qualitative analysis that involve semi structured individual interview.

Figure 1: Research Process

Figure 1 illustrates a step-by-step process for conducting qualitative research using semi-structured interviews and thematic analysis. The process begins with formulating clear research questions, which guide the development of an interview guide containing open-ended questions. Suitable participants are then recruited. Semi-structured interviews are conducted, allowing for in-depth exploration of participants' experiences. The audio recordings from these interviews are transcribed verbatim to ensure accurate data capture. The transcribed data is then analyzed using thematic analysis, where patterns and themes are identified, coded, and interpreted. Finally, the findings are interpreted in relation to the research questions and reported in a structured format to provide meaningful insights.

As a preliminary study, a total of six participants were interviewed in-depth to explore key issues and identify initial themes. This number is sufficient for early exploratory purposes and is supported by previous studies (Creswell, 2012). Students who have taken Introductory Arabic Language (level 1) from different faculty in UiTM Shah Alam were selected to be interviewed and some lecturers have been asked for their students' voluntary participation.

McDonough and McDonough (1997) pointed out that it is natural for both researchers and informants to use the language of their mother tongue. Thus, the informants were allowed to use both English and Bahasa Melayu to warrant that the informants felt at ease as it was easier for them to share their views and perception. Adapted from Norhaslinda (2022), this study used

e-ISBN: 978-629-98082-2-0

semi structured individual interview questions as a research instrument in which the interview questions are based on the research questions.

3. Results and Discussion

This study explored the washback effects of the Introductory Arabic Language Writing Assessment (IALWA) on student learning. Findings revealed a spectrum of student perceptions regarding the assessment's difficulty. Those with prior Arabic language exposure found the test relatively manageable, while students without foundational knowledge expressed challenges, especially in applying grammar rules and recognizing patterns. These varied perceptions directly influenced motivation: students who found the assessment achievable were more engaged, suggesting a positive motivational washback. On the other hand, those who struggled expressed anxiety and reduced confidence, illustrating the potential for negative washback in the absence of sufficient scaffolding.

Technical and logistical barriers also emerged as notable influences. Many students cited slow internet connections and difficulty using the Arabic keyboard as obstruction during the test. These issues, unrelated to their actual language ability, affected their performance and introduced a dimension of stress unrelated to the learning objectives. This highlights how non-pedagogical factors can interfere with intended assessment outcomes and diminish constructive washback.

Despite these barriers, the writing assessment was largely recognized as valuable for reinforcing learning. Students indicated that the test encouraged them to solidify their vocabulary and grammar knowledge and apply them more confidently. This indicates that IALWA can support constructive washback by aligning assessment content with key learning outcomes.

Lecturers played a crucial role in shaping student experiences. Most participants praised their lecturers' clear instruction, structured revision strategies, and frequent use of practice exercises. Such teaching practices enhanced student understanding and contributed to a positive learning environment, reinforcing the benefits of assessment-aligned instruction. Dummy tests, in particular, served as effective preparation tools, enabling students to familiarize themselves with assessment formats and identify gaps in knowledge prior to the actual test.

Student learning extended beyond the classroom. Many engaged in peer revision, memorization strategies, and textbook exercises. The assessment thus motivated independent and collaborative learning, further demonstrating a washback loop where assessment requirements encouraged more comprehensive engagement with the course content.

e-ISBN: 978-629-98082-2-0

Table 1: Thematic Overview of Students' Experiences with Arabic Writing Assessment

Theme	Frequency	Response (English Translation)
Perceived Difficulty of the Writing Assessment	6	It's not difficult if you study well and focus in class.
		It wasn't easy but it wasn't too difficult either.
		Without revision, students may struggle.
		It's necessary and compulsory, but difficult due to digital format.
		Arabic isn't that hard... some parts are familiar from school.
		At level one, it's easy for me because I have Arabic basics.
Technical and Format Challenges	5	The WiFi was slow and my hotspot was also slow.
		The Arabic keyboard can be a bit difficult to use.
		Time to submit was limited due to poor connection.
		Arabic grammar is tricky with added letters and endings.
		A bit hard to type using the Arabic keyboard.
Importance of Writing Assessment	5	It's important because we use the Arabic keyboard.
		We must learn how to write the language to better understand it.
		To evaluate the student's understanding of the subject.
		We can tell if students understand what was taught.
		I think it's important because it can test the students' understanding.
Effective Teaching Methods	5	The lecturer gave a dummy and explained how to answer.
		I didn't get any misinformation about the test.
		The lecturer gave step-by-step answering tips.
		The way the lecturer teaches is not boring and very clear.
		My lecturer explains everything in detail, step-by-step.
In-Class Learning Strategies	4	I revised with the lecturer in class.
		I did all the exercises in the textbook.
		The lecturer did a quick revision in class.

e-ISBN: 978-629-98082-2-0

Theme	Frequency	Response (English Translation)
		I wrote the meanings and reviewed them before the test.
Out-of-Class Learning Strategies	5	I revised with my friends.
		The three of us practiced and revised together.
		We chatted in Arabic to help remember words.
		I read the book and asked a friend if I didn't understand.
		We formed study groups and helped each other revise.
Impact of Dummy Test and Revision Materials	5	The dummy test really helped as early preparation.
		The test file helped us practice for the real one.
		The lecturer gave a dummy for students to try and see.
		I finished all the learning before doing the dummy test.
Students' Emotional and Motivational Responses	4	I like learning Arabic because it's fun.
		I like my lecturer this semester.
		I like to attend this Arabic class; easy to understand.
		I like it because I have a basic background in Arabic.

In summary, this study identified both positive and negative washback effects. Positive effects were associated with clear teaching practices, formative tools like dummy tests, and student motivation. On the other hand, negative effects stemmed from technical limitations and students' lack of prior knowledge. These findings emphasize the critical need for assessments to be both pedagogically valid and practically accessible in terms of their format and implementation.

4. Conclusion

In conclusion, the Introductory Arabic Language Writing Assessment (IALWA) demonstrates a significant influence on student learning, functioning as both a motivator and a challenge. While it successfully encouraged vocabulary acquisition, grammar application, and active revision strategies, its effectiveness was moderated by students' prior exposure to Arabic and technical barriers during administration. To maximize positive washback, educators and curriculum designers should ensure that assessment tasks are clearly communicated, aligned with

e-ISBN: 978-629-98082-2-0

instructional content, and supported by robust preparatory activities such as practice tests, thorough review, seeking clarification and creating study aids; for example developing flashcards and concept maps and notes. Moreover, logistical challenges—particularly in digital settings—should be minimized to prevent external factors from undermining assessment reliability. This study highlights the necessity for thoughtfully designed assessments that foster deeper engagement with language learning while remaining sensitive to students' diverse needs and contexts.

Acknowledgement

This work was supported by Geran Inisiatif Akademi Pengajian Bahasa (GIA).

References

Alderson, J.C. & Wall, D. (1993). Does Washback Exist? Applied Linguistics. 14(2): 115-129. DOI:10.1093/APPLIN/14.2.115.

Bachman, L.F., Palmer A.S. (1996). Language Testing in Practice: Designing and Developing Useful Language Test. Oxford. Oxford University Press.

Brown, J. (2004). University Entrance Examinations: Strategies for Creating Positive Washback on English language Teaching in Japan. . Shiken: JALT Testing and Evaluation SIG Newsletter Vol 3 (2). p. 4-8. Retrieved from http://www.jalt.org/test/bro_5.htm.

Creswell, J. W. (2012). Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research, 4th ed. Boston, MA: Pearson Education.

Gipps, C.V. (1994). Beyond Testing: Towards a Theory of Educational Assessment. London: Routledge. Retrieved from: https://www.researchgate.net/publication/360835676_Beyond_Testing_Towards_a_Theory_of_Educational_Assessment_by_Caroline_V_Gipps.

Hossain, Md. Mahroof, Ahmed, Md. Kawser. (2015). Language Testing: An Overview and Language Testing in Educational Institutions of Bangladesh. Advances in Language and Literary Studies. Vol 6 (6), 80-84. http://dx.doi.org/10.7575/aiac.alls.v.6n.6p.80.

Kolahi, Sholeh, 2007. Investigating The Washback Effects On Improving The Writing Performance Of Iranian EFL University Students. Faculty of Communication and Modern Languages, Universiti Utara Malaysia. Retrieved from https://repo.uum.edu.my/id/eprint/3282/1/Sho.pdf

McDonough, J., & McDonough, S. (1997). Research methods for English language teachers. London, England: Arnold.

Nopita., D., 2019. Washback in EFL Writing Courses: A Reflective Teaching at English Education Study Program of Teacher Training and Education Faculty of Universitas

e-ISBN: 978-629-98082-2-0

Maritim Raja Ali Haji. Journal of Language Learning and Research, 2(1), 1-11. https://doi.org/10.22236/jollar.v2i1.3486

Norhaslinda, H. (2022). Washback Study of an Outcome Based English Language Assessment on Student Learning. (Doctoral dissertation, International Islamic University Malaysia). https://hikmahlib.iium.edu.my/cgi-bin/koha/opac-detail.pl?biblionumber=514325.

Pearson, L. (1988). Tests as levers of change (or "putting first things first"). In D. Chamberlain & R. Baumgartner (Eds.), ESP in the classroom: Practice and evaluation ELT Documents #128, (pp. 98-107), Modern English Publication in association with the British Council, London

Tayeb, et al., 2014. The washback effect of the general secondary english examination (GSEE) on teaching and learning. GEMA Online Journal of Language Studies, Vol. 14(3). Accessed from http://dx.doi.org/10.17576/GEMA-2014-1403-06

e-ISBN: 978-629-98082-2-0

P080

Can AI Replace Human Raters? A Multi-Dimensional Analysis of AES Reliability Using GPT-4 and Beyond

***Dai yi[1], Du Meirong[1], Wang Fan[2], Lu Min[3], Zhang Yu[4]**

*[*1,1,4]School of Education, City University of Macau, Macau 999078, China.*
[2-3]University International College, Macau University of Science and Technology, Macau 999078, CHINA.
*(E-mail:[1]M24092100349@cityu.edu.mo, [2]2240012955@student.must.edu.mo, [3]2240027464@student.must.edu.mo, [4]*yzhangin06@outlook.com*)*
**corresponding author: [1]yidai@cityu.edu.mo*

Abstract

The study investigates the potential and limitations of AI tools in high-stakes writing assessment, particularly in the IELTS context. Thirty writing samples from IELTS Skills course students were evaluated by two human raters and three AI tools (ChatGPT, DeepSeek, and Kimi) using the official IELTS writing rubric. Results showed significant differences between AI tools and human raters in overall writing scores ($F = 5.319$, $p < .001$, $\eta^2 = 0.314$), with human raters outperforming AI tools. While AI tools demonstrated comparable performance in assessing task response (TR), significant discrepancies were found in coherence and cohesion (CC), lexical resource (LR), and grammatical range and accuracy (GRA). ChatGPT showed the highest consistency with human raters in assessing GRA. The findings suggest that AI tools can provide supplementary feedback in specific areas but cannot fully replace human raters due to their limitations in evaluating higher-order writing skills. Future research should explore larger sample sizes and diverse AI tools to further validate these findings.

Keywords: Automated Writing Evaluation (AWE); AI tools; IELTS writing test; human raters; scoring consistency

1. Introduction

The rapid advancement of artificial intelligence (AI) has revolutionized various fields, including education, particularly in the area of automated writing evaluation (AWE). AWE systems have evolved significantly over the past decades, from early attempts in the 1960s to the sophisticated models of today. The development of AWE has progressed through three key phases. Early systems, such as Page's (1966) Project Essay Grader, relied on superficial features like word count and essay length, but failed to assess the quality of writing (Dikli, 2006). The second generation of AWE tools, such as the e-rater by the Educational Testing Service (ETS),

e-ISBN: 978-629-98082-2-0

introduced natural language processing (NLP) techniques to analyze lexical complexity and basic coherence (Mizumoto & Eguchi, 2023). However, these tools were still limited in their ability to evaluate rhetorical sophistication and higher-order writing skills (Khalifa & Albadawy, 2024). The contemporary generation of AWE systems, based on advanced transformer architectures like GPT-4, demonstrates significantly improved contextual understanding and the ability to provide feedback on higher-order writing elements (Dai et al., 2023). Despite these advancements, gaps exist between AI-generated scores and human raters' nuanced judgments, especially in high-stakes assessment contexts (Latif & Zhai, 2024).

While traditional AWE systems excelled in assessing surface-level features such as grammatical accuracy and lexical richness (Attali & Burstein, 2006), their inability to evaluate rhetorical sophistication and coherence created a 'coherence gap' (Saralajew et al., 2022). Recent studies have shown that even the most advanced AI tools, such as ChatGPT, still demonstrate systematic overestimation of coherence by up to 0.5 band scores compared to human raters (Wetzler et al., 2024). Additionally, feedback depth limitations persist, as evidenced by Grammarly missing 62% of lexical errors identified by human experts (Park, 2019). Cultural bias is another critical issue, with AI tools often penalizing expressions influenced by the first language (L1) of non-native speakers (Almegren et al., 2024).

Despite these challenges, the potential of AI in writing evaluation remains promising. AI tools have the capacity to provide instant and consistent feedback, significantly reducing the workload of human teachers (Han & Li, 2024). They can analyze a wide range of writing features and offer detailed feedback that can help students improve their writing skills (Khalifa & Albadawy, 2024). However, the effectiveness of AI tools in replacing human raters' feedback and scoring remains an ongoing subject of research (Lo et al., 2025). This study aims to explore the potential and limitations of AI tools in high-stakes writing assessment, particularly in the context of the IELTS writing test. By comparing the scoring consistency of multiple AI tools (ChatGPT, DeepSeek, and Kimi) with human raters across different writing dimensions, this research seeks to address the critical question of whether these prevalent AI tools can effectively bridge the human-AI scoring divide in high-stakes writing assessment. The findings of this study have significant implications for both educational measurement theory and the practical implementation of AI-assisted evaluation systems.

2. Methods

This study aimed to evaluate the performance of three AI tools—ChatGPT, DeepSeek, and Kimi—against human raters in assessing IELTS writing tasks. Thirty writing samples were collected from students enrolled in an IELTS Skills course in southern China. These students

e-ISBN: 978-629-98082-2-0

were targeting a band score of 6.5 or higher, indicating their proficiency level and preparation for the IELTS exam. The samples were selected to ensure homogeneity and relevance to the study's objectives.

The assessment criteria were based on the official IELTS writing rubric, which evaluates four key components: Task Response (TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA). Each component was scored out of 9.0, with the overall score calculated as the average of the four component scores. This rubric is widely used in IELTS assessments and provides a standardized framework for evaluating writing proficiency.

Two IELTS-certified human raters independently scored the papers, ensuring that the scoring process was unbiased and consistent. The average score from these two raters was recorded for each paper. The same prompts were used for the AI tools (ChatGPT, DeepSeek, and Kimi) to ensure consistency in the evaluation process. Each writing sample was evaluated separately by the AI tools and human raters, allowing for a direct comparison of their assessments. To analyze the data, the scores from the AI tools and human raters were compared using one-way ANOVA. This statistical method was chosen to determine significant differences in mean scores across different dimensions. Effect sizes were calculated using eta-squared (η^2) to measure the extent of these differences.

3. Results and Discussion

This study aimed to explore the potential and limitations of AI tools in high-stakes writing assessment by comparing their performance with human raters using the IELTS writing rubric. The results (Table 1) revealed significant differences between AI tools and human raters in overall writing scores ($F = 5.319$, $p < .001$, $\eta^2 = 0.314$), with human raters consistently outperforming AI tools. Among the AI tools, Kimi achieved the highest mean score ($M = 5.82$, $SD = 0.09$), followed by DeepSeek ($M = 5.48$, $SD = 0.16$) and ChatGPT ($M = 5.33$, $SD = 0.14$). However, the human examiner group attained the highest mean score ($M = 5.98$, $SD = 0.13$). The 95% confidence intervals further supported these findings, with the human examiner's lower limit ($LL = 5.61$) exceeding the upper limits of ChatGPT ($UL = 5.70$) and DeepSeek ($UL = 5.55$).

Table 1: Overall writing scores by the human rater and AI tools

				95% Confidence Interval for Mean				
		Mean	Std. Deviation	LL	UL	F	Sig.	η^2
Overall	DeepSeek	5.48	0.159	5.81	6.55	5.319	<.001	0.314

e-ISBN: 978-629-98082-2-0

Kimi	5.82	0.093	5.45	6.19
ChatGPT	5.33	0.136	4.96	5.70
Human Examiner	5.98	0.126	5.61	6.35

In terms of specific dimensions, AI tools demonstrated comparable performance to human raters in assessing Task Response (TR), with no statistically significant differences ($F = 0.894$, $p = 0.452$, $\eta^2 = 0.026$) (Table 2). However, significant discrepancies emerged in Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA). For CC, DeepSeek scored higher than human raters ($M = 6.22$, $SD = 0.20$), while ChatGPT ($M = 5.58$, $SD = 0.18$) and Kimi ($M = 5.88$, $SD = 0.11$) scored lower. In LR, DeepSeek again surpassed the human benchmark ($M = 6.19$, $SD = 0.23$), while ChatGPT ($M = 5.57$, $SD = 0.22$) and Kimi ($M = 5.63$, $SD = 0.13$) scored lower. For GRA, ChatGPT showed the highest consistency with human raters ($M = 5.65$, $SD = 0.10$), while DeepSeek ($M = 5.93$, $SD = 0.15$) overscored and Kimi ($M = 5.40$, $SD = 0.15$) underscored.

Table 2. The quality of assessment by the human rater and AI tools

				95% Confidence Interval for Mean				
		Mean	Std. Deviation	LL	UL	F	Sig.	η^2
TR	DeepSeek	5.93	0.145	5.58	6.28	0.894	0.452	0.026
	Kimi	5.83	0.145	5.48	6.18			
	ChatGPT	5.98	0.136	5.63	6.33			
	Human Examiner	6.03	0.176	5.68	6.38			
CC	DeepSeek	6.22	0.203	5.85	6.58	8.957	<.001	0.571
	Kimi	5.88	0.109	5.52	6.25			
	ChatGPT	5.58	0.183	5.22	5.95			
	Human Examiner	6.03	0.120	5.67	6.40			
LR	DeepSeek	6.19	0.233	5.78	6.58	7.342	<.001	0.305
	Kimi	5.63	0.133	5.23	6.03			
	ChatGPT	5.57	0.219	5.17	5.97			
	Human Examiner	6.00	0.115	5.60	6.40			

e-ISBN: 978-629-98082-2-0

GRA	DeepSeek	5.93	0.145	5.68	6.19	12.342	<.001	0.637
	Kimi	5.40	0.153	5.15	5.65			
	ChatGPT	5.65	0.100	5.41	5.92			
	Human Examiner	5.70	0.058	5.45	5.95			

This study explores the use of AI tools in high-stakes writing assessment, particularly in the context of the IELTS writing test. The results reveal both promising consistencies and notable discrepancies between AI-generated scores and those assigned by human raters. While AI tools demonstrated comparable performance to human raters in assessing Task Response (TR), significant differences were found in Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA). These findings highlight the limitations of AI tools in evaluating higher-order writing skills and the nuanced judgments of human raters.

Human raters consistently outperformed AI tools in overall scores and most categories, demonstrating a more nuanced understanding of essay evaluation. This aligns with previous studies that highlight the limitations of AI tools in Automated Essay Scoring (AES) (Manning et al., 2025; Almegren et al., 2024). However, AI tools showed potential in specific areas, such as providing instant feedback and reducing the workload of human teachers (Han & Li, 2024).

The high level of agreement between AI tools and human raters in assessing TR suggests that AI can effectively determine whether students have adequately addressed the writing prompt. However, discrepancies in CC and LR indicate that AI tools may struggle with more nuanced aspects of writing, such as logical flow and vocabulary complexity. ChatGPT showed the highest consistency with human raters in assessing GRA, suggesting it may be a useful tool for evaluating grammatical accuracy (Latif & Zhai, 2024).

Despite these findings, the study highlights the importance of human raters in providing context-sensitive and nuanced feedback. AI tools should be used as supplementary tools rather than replacements for human raters. Educators can leverage AI for initial scoring and specific feedback, while human raters focus on higher-order feedback. Future research should explore larger sample sizes, diverse AI tools, and the potential for combining AI with human raters to improve writing assessment.

This study provides insights into AI tools for writing assessment but has limitations. The small sample size of 30 IELTS students from southern China restricts generalizability. Future research should use larger, more diverse samples and include additional AI tools like Gemini for a

e-ISBN: 978-629-98082-2-0

comprehensive understanding. Testing applicability in other high-stakes exams like TOEFL is also needed. Developing advanced AI models to evaluate complex writing skills and exploring AI-human rater collaboration to enhance assessment quality are recommended.

4. Conclusion

This study provides valuable insights into the potential and limitations of AI tools in high-stakes writing assessment by comparing their performance with human raters. The findings reveal that while AI tools demonstrate comparable scoring abilities to human raters in certain dimensions, significant discrepancies remain in overall scoring and the evaluation of key writing features. Specifically, human raters excel in assessing the overall quality of writing, depth of content, and lexical resources, whereas AI tools show higher consistency in evaluating grammatical range and accuracy. These results underscore the nuanced and context-sensitive feedback that human raters can provide, as well as the specific strengths of AI tools in certain assessment tasks.

In practical terms, educators are encouraged to adopt a tiered feedback model, involving AI tools for initial scoring and human raters for higher-order feedback. This approach can significantly reduce the workload of educators while ensuring that students receive comprehensive feedback. Future research should explore larger sample sizes and diverse AI tools to further validate these findings and improve AI's role in writing assessment.

Acknowledgement

This work was supported by the the Macao Science and Technology Development Fund (FDCT) under Grant [number 0071/2023/RIB3]; the Joint Research Funding Program between the Macau Science and Technology Development Fund (FDCT) and the Department of Science and Technology of Guangdong Province (2024) (FDCT-GDST) under Grant [number 0003-2024-AGJ].

References

Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2024). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1–16. https://doi.org/10.1080/14703297.2024.2437122

e-ISBN: 978-629-98082-2-0

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, 4*(3), 1-30 http://download.chasedream.com/gmat/awa/Automated_Essay_Scoring_v2.pdf

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE international conference on advanced learning technologies (ICALT)* (pp. 323-325). IEEE.. https://doi.org/10.35542/osf.io/hcgzj

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1-35. http://files.eric.ed.gov/fulltext/EJ843855.pdf

Han, J., & Li, N. M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System, 126*, 103502. https://doi.org/10.1016/j.system.2024.103502

Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update, 5*, 100145. https://doi.org/10.1016/j.cmpbup.2024.100145

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lo, N., Wong, A., & Chan, S. (2025). The impact of generative AI on essay revisions and student engagement. *Computers and Education Open,* 100249. https://doi.org/10.1016/j.caeo.2025.100249

Manning, J., Baldwin, J., & Powell, N. (2025). Human versus machine: The effectiveness of ChatGPT in automated essay scoring. *Innovations in Education and Teaching International*, 1–14. https://doi.org/10.1080/14703297.2025.2469089

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), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Park, J. (2019). An AI-based English grammar checker vs. human raters in evaluating EFL learners' writing. *Multimedia-Assisted Language Learning, 22*(1), 112–131. https://doi.org/10.15702/mall.2019.22.1.112

Saralajew, S., Shaker, A., Xu, Z., Gashteovski, K., Kotnis, B., Rim, W. B., Quittek, J., & Lawrence, C. (2022). A Human-Centric assessment framework for AI. *ArXiv Preprint ArXiv:2205.12749.* https://doi.org/10.48550/arxiv.2205.12749

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2024). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology, 52*(3), 298-304. https://doi.org/10.1177/00986283241282696

e-ISBN: 978-629-98082-2-0

P087

Assessment Tasks and Autonomous Learning Motivation in Secondary Vocational English Classrooms: An Analysis Based on Structural Equation Modeling

***He Yang[1], Sihui Yu [2], Yiran Miao [3]**

[1]School of Education, City University of Macau, Macau. [2]Mental Health Education and Counseling Center, Shenzhen Polytechnic University, China. [3] College of Foreign Languages, Chengdu University of Information Technology, CHINA.
(E-mail: [1]M23092100755@cityu.edu.mo, [2]yusihui693380@126.com, [3]myr@cuit.edu.cn)
**Corresponding author: [1]M23092100755@cityu.edu.mo*

Abstract

Within China's integrated teaching-learning-assessment framework, vocational students exhibit severe autonomy deficits in English learning (e.g., low interaction/passive engagement). The perception-motivation linkage remains underexplored in vocational contexts. To examine predictive pathways of assessment task perceptions (Congruence, Authenticity, Consultation, Transparency, Diversity) on autonomous motivation based on Brookhart's (1997) classroom assessment theory. Adapted PATI and ALM scales were administered to 281 vocational students from a nationally recognized Shanghai institution (response rate: 89.8%). Structural equation modeling (SEM) was employed with fit indices meeting CFI≥0.90, SRMR/RMSEA<0.08 thresholds. Congruence (β=0.32), authenticity (β=0.41), and diversity (β=0.27) significantly predicted autonomous motivation, whereas transparency and consultation showed non-significant effects ($p > .05$). This study validates the cross-cultural applicability of Brookhart's tri-dimensional motivators in vocational education, demonstrating task-driven motivation through competence-need fulfillment (authenticity→efficacy→motivation), and informs pedagogies for tiered authentic task design.

Keywords: assessment tasks, autonomous learning motivation, students' perception, secondary vocational school students

1. Introduction

Within China's integrated teaching-learning-assessment framework, classroom assessment's role in promoting learning has gained significant attention. Some empirical studies have identified significant learning motivation deficits among secondary vocational school students (Hao & Pilz, 2021; Liu, 2024). This study examines how English classroom assessment tasks

e-ISBN: 978-629-98082-2-0

(Brookhart, 2014) impact vocational students' autonomous motivation. Students' perceptions of assessment environments influence learning behaviors through motivational mediation (Rabari & Indoshi, 2011). Therefore, we investigate the relationship between task perception and autonomous motivation from learners' perspectives, aiming to ignite intrinsic drive and optimize learning outcomes.

Students' perceptions of assessment tasks—regarding importance, value, difficulty, and success likelihood—shape motivational beliefs (McMillan & Workman, 1998). Brookhart's (1997) classroom assessment model identifies five foundations: teacher attitudes, assessment purposes, preparedness, instructional integration, and feedback. Perceptions formed through collective assessment experiences directly impact motivation and achievement (Brookhart, 2014). Task value/difficulty/success-probability cognitions during engagement critically influence motivation (McMillan & Workman, 1998). Self-efficacy is modulated by task characteristics: excessive difficulty or unclear criteria weaken efficacy (McMillan & Workman, 1998), whereas authenticity (Van Dinther et al., 2014), perceived meaningfulness, capability alignment, and private feedback (Alkharusi & Al-Hosni, 2015) enhance efficacy judgments and deep learning. High task authenticity/transparency particularly bolsters self-efficacy in science (Alkharusi, 2013). Dorman & Knightley's (2006) five-dimensional PATI scale (Congruence, Authenticity, Consultation, Transparency, Diversity), globally validated across contexts (e.g., Zhang & Li, 2023), is the dominant tool for measuring task perceptions. This study applies it to test Brookhart's model among Shanghai vocational students in China.

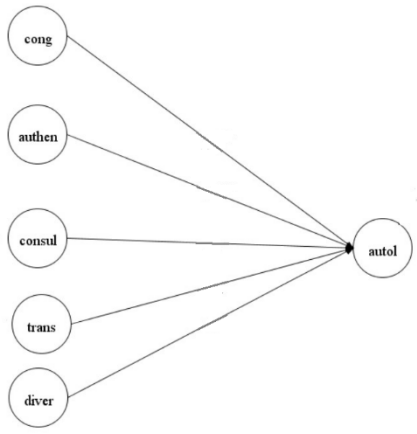
Brookhart (1997) establishes four motivational dimensions of task perception: inappropriate difficulty reduces effort/achievement, perceived importance drives engagement, interest triggers intrinsic motivation, and value recognition enhances persistence. A critical research gap persists in the classroom assessment perception-autonomous motivation nexus—particularly significant as motivation types profoundly impact academic outcomes (Zimmerman, 2001) and learning strategies (Ng et al., 2016). Self-Determination Theory (SDT) provides the key framework: assessment environments must fulfill autonomy (value-driven agency), competence (mastery confidence), and relatedness (respect-based bonds) (Ryan & Deci, 2017). Autonomous motivation (intrinsic/identified/integrated regulation) fundamentally differs from controlled motivation: the former stems from integrated self-concepts (Deci et al., 2020), the latter from external pressure (Mouratidis et al., 2021). Assessment tasks exert dual effects: exams may promote career-goal-oriented learning (Li & Yang, 2023) yet simultaneously undermine intrinsic motivation through grading pressure (Pulfrey et al., 2011). Current measurement adopts dual approaches: direct scales (e.g., Macaskill and Taylor's (2010) Learning Independence/Habits bifactor) or indirect indices (e.g., Vallerand's (1997) Relative Autonomy Index). Nevertheless, this pivotal controversy remains underexplored in classroom

e-ISBN: 978-629-98082-2-0

assessment theory.

In China, there is a scarcity of high-quality research on the relationship between assessment and autonomous learning motivation, as well as academic achievement in secondary vocational English classrooms. Chinese research on English classroom assessment seldom covers the secondary school level compared to international research (Jin & Sun, 2020), and even less so for secondary vocational English classrooms. So, this study aims to build and test an analytical model explaining how secondary vocational students' perceptions of English classroom assessment task characteristics affect autonomous learning motivation. The research questions are: (1) What are the characteristics of assessment tasks as perceived by students in secondary vocational English classrooms? (2) Do these perceived assessment task features influence students' autonomous learning motivation? Following previous research showing positive effects of assessment task perceptions on motivational variables (Alkharusi et al., 2013b), this study proposes a theoretical model (see Figure 1) hypothesizing that students' views on assessment tasks in terms of congruence with planned learning, authenticity, student consultation, transparency, and diversity will directly positively impact autonomous learning motivation.

Figure 1: Model of the relationship between students' perceptions of assessment tasks and autonomous learning motivation



2. Methods

2.1 Participants

A convenience sample of 313 vocational students from a nationally recognized Shanghai institution participated in school-level English open classes. Open classes ensured rigorous assessment task design (valid questionnaires: 281; response rate: 89.8%). Demographics included: gender ratio (55% female), disciplinary distribution (37.4% STEM majors), program

e-ISBN: 978-629-98082-2-0

composition (71.2% integrated secondary-vocational track; 28.8% regular vocational track), and balanced self-reported academic ranking (top/middle/bottom = 1:1:1).

2.2 Procedures

Data collection was conducted during the class meeting that followed the open English classes on the same day, each student received a Chinese self-report questionnaire consisting of three parts. The participating students were informed that they were taking part in a study investigating their views on classroom assessment tasks; participation in the questionnaire was not mandatory; and the information they provided would be kept confidential. The two parts of the questionnaire were as follows: the first part contained demographic information about the students, including their major, class type, and self-reported academic ranking within their class. The second part included measurements of students' perceptions of assessment tasks and autonomous learning motivation. The questionnaire required students to consider their perceptions in the completed open English class course when evaluating these perceptions. Under the guidance of the English teacher or class teacher, students needed to spend about 20 minutes completing the electronic questionnaire using their mobile phones.

2.3 Instruments

Two core constructs were measured using 7-point Likert scales (1=strongly disagree, 7=strongly agree): students' perception of assessment tasks was adapted from Dorman and Knightley's (2006) five-dimensional PATI, comprising 35 items (e.g., "English assessment tasks reflect real-world contexts"); autonomous learning motivation (ALM) derived from Macaskill and Taylor's (2010) Learning Independence subscale, refined to 7 items (e.g., "I proactively seek new learning experiences"). Dimension scores were calculated as item means, with higher values indicating more positive perceptions.

2.4 Data Analysis

Data analysis was conducted using SPSS 26 and Mplus 8. After screening 313 questionnaires for missing values, 32 invalid responses were excluded, retaining 281 valid cases. Exploratory factor analysis established construct validity: principal axis factoring with oblique rotation for the assessment task perception scale, and principal component analysis extracting a single factor for the autonomous motivation scale. Confirmatory factor analysis then verified the factor structure (fit indices: CFI≥0.90, SRMR/RMSEA<0.08), with Cronbach's α assessing internal consistency reliability. Descriptive statistics computed means, standard deviations, and Pearson correlations. Structural equation modeling ultimately examined the perception-motivation pathway, with model fit evaluated against Hu and Bentler's (1999) composite criteria (CFI/RMSEA/SRMR). A CFI of 0.90 indicated that the model fit was reasonable; RMSEA and SRMR both less than 0.08 indicated that the model fit was reasonable (Hu &

e-ISBN: 978-629-98082-2-0

Bentler, 1999); however, if RMSEA was less than 0.1, the model fit was also considered acceptable (Browne & Cudeck, 1993).

3. Results and Discussion

3.1 Exploratory Factor Analysis (EFA) and Internal Consistency Analysis

Principal axis factoring with oblique rotation was employed for the assessment task perception scale. Items with factor loadings <0.4 or cross-loadings >0.3 were eliminated (affecting Authenticity, Consultation, Transparency, and Diversity subscales), resulting in a five-factor structure with 25 retained items. Each dimension contained ≥4 items (see Table 1), collectively explaining 82.9% of variance. The scale demonstrated excellent internal consistency (total α=0.94; subscale α=0.86-0.97), confirming robust construct validity and reliability.

Table 1: EFA Results of The Perception of Assessment Tasks

	Factor					Cronbach α
	Congruence	Transparency	Consultation	Diversity	Authenticity	.94
Congruence1-7	.60-.91					.95
Transparency 1-4		.61-.97				.94
Consultation 1-4			-.93- -.65			.92
Diversity1-4				.53-.84		.86
Authenticity 3					-.97- -.57	.97

For the autonomous learning motivation scale, since there was only one-dimension, principal component analysis was used to extract one factor. The factor loadings for each item ranged from 0.49 to 0.95, with a cumulative variance explained of 78.5%. The Cronbach's alpha value was 0.93, indicating internal consistency.

3.2 Confirmatory Factor Analysis (CFA)

The structures of the Perceived Assessment Task Inventory (PATI) consisting of five factors and the Autonomous Learning Motivation scale (ALM) consisting of one factor, as derived from the EFA, were analyzed using CFA. The initial fit indices for the two CFA models were not ideal, so revisions were made to the models based on modification indices (MIs) and literature. The fit indices of the revised models were acceptable. Table 2 lists the specific fit index values calculated for the scales by CFA.

e-ISBN: 978-629-98082-2-0

Table 2: The CFA Findings of PATI and ALM

Model	chi-square	df	CFI	SRMR	RMSEA (90% CI)
PATI (original)	1022.138***	265	.915	.079	. 101**(.094,.107)
PATI (revised)	719.084**	260	.948	.075	.079**(.072,.086)
ALM (original)	401.919**	14	.866	.050	.314**(.288,.341)
ALM (revised)	37.063**	10	.991	.037	.098**(.066,.133)

*Note. **p<0.01. PATI =Assessment task perception Model Assessment task, ALM = Autonomous learning motivation.*

3.3 Descriptive Statistical Results

Descriptive statistics (Table 3) revealed participants held the most positive perceptions of task authenticity (highest M), followed by congruence with planned learning. All five assessment task characteristics received favorable evaluations in English open classes, with generally high autonomous motivation levels. Correlation analysis indicated statistically significant positive interrelationships among task characteristics, showing strong congruence-authenticity associations (highest r) but weak transparency linkages. Task perceptions collectively correlated positively with autonomous motivation, demonstrating particularly strong relationships with congruence and authenticity yet weaker ties to transparency.

Table 3: Descriptive Statistical Results

	cong	auth	cons	tran	dive	M	SD.
cong	-					6.35	.98
auth	.80**	-				6.45	.92
cons	.49**	.48**	-			5.57	1.63
tran	.18**	.14*	.51**	-		4.67	2.06
dive	.35**	.31**	.69**	.88**	-	5.31	1.64
AutoL	.76**	.78**	.43**	.18**	.34**	6.28	.99

*Note. *p<0.05, **p<0.01. cong=congruence with planned learning, auth=Authenticity, cons= Student consultation, tran=Transparency, dive=Diversity.*

3.4 SEM Analysis Results

Based on Table 4, the fit indices of the initial structural equation model were not ideal. The model was refined by combining theoretical analysis and modification indices, resulting in a revised structural equation model with acceptable fit. The specific indices of the initial and revised models are presented in the table below. According to the SEM results (see Chart 2), the students' perception of assessment tasks can partially predict autonomous learning motivation. Specifically, congruence with planned learning, authenticity, and diversity have significant predictive effects on autonomous learning motivation, while transparency and student consultation do not significantly predict autonomous learning motivation.

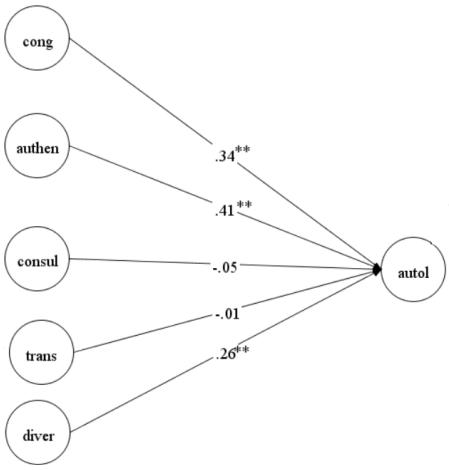
e-ISBN: 978-629-98082-2-0

Table 4: The Results of SEM

Model	chi-square	df	CFI	SRMR	RMSEA（90% CI）
SEM (original)	1716.917**	449	0.897	0.072	0.100**（0.095，0.105）
SEM (revised)	1046.964**	441	0.951	0.068	0.070**（0.064，0.075）

*Note. **p<0.01.*

Chart 2: Revised SEM results of the relationship between perceived assessment tasks and autonomous learning motivation



*Note. **p<0.01.*

3.5 Discussion

3.5.1 Regarding the Partial Predictive Role of Assessment Task Perception on Autonomous Learning Motivation

In secondary vocational English classrooms, students' perceived assessment task authenticity most strongly predicts autonomous learning motivation, followed by congruence with planned learning, with diversity having the weakest predictive power, partially supporting the hypotheses.

Authentic assessment tasks boost learning motivation (Frey, Schmitt & Allen, 2012), with this study showing authenticity can partially predict autonomous motivation. McMillan and Workman (1998) noted that higher authenticity links assessment to real-life activities, enhancing its significance, practicality, and value, thus strengthening motivation. Exams with authentic tasks integrate assessment with daily activities promoting autonomous learning (Zhang & Li, 2019). Students perceiving higher authenticity are more likely to develop learning ability and positive beliefs about materials (Alkharusi, 2013).

e-ISBN: 978-629-98082-2-0

As congruence with planned learning enhances academic self-efficacy (Alkharusi et al., 2013b, 2014), which affects motivation (Pajares & Schunk, 2001), it can influence autonomous motivation. Lizzio and Wilson (2013) suggest that when students see assessment tasks as aligned with goals and valuable, they want and believe they can learn well, leading to proactive learning and better academic performance through deep-learning strategies (Gulikers et al, 2008).

Like congruence, diversity's impact on motivation may stem from its indirect positive effect on self-efficacy (Alkharusi et al., 2013b, 2014). However, diversity doesn't necessarily improve academic performance. Alkharusi et al. (2013a) found no link between student grades and diversity, implying a motivated environment from diverse tasks doesn't guarantee better academic outcomes due to other influencing factors.

3.5.2 Student consultation and transparency did not predict autonomous learning motivation

In secondary vocational English classrooms, students' perceived student consultation and transparency of assessment tasks weren't significant predictors of autonomous learning motivation, conflicting with some research hypotheses. This paper explains this from the literature and current assessment practices in these classrooms.

Student consultation necessitates teachers soliciting feedback on assessment formats for pedagogical improvement. However, vocational students' foundational deficiencies and motivational deficits (Chuane et al., 2023) manifest as classroom disengagement (e.g., sleeping, phone use), reinforcing teacher dominance and distrust in student input. Self-Determination Theory (Deci & Ryan, 2000) identifies competence needs (self-efficacy) as pivotal for autonomous motivation: only when students perceive capability can autonomous drive be ignited (Joe et al., 2017). Empirical evidence confirms consultation's non-significant impact on efficacy (Alkharusi et al., 2013b), disabling its motivational function through the competence-need pathway and severing the "consultation-efficacy-motivation" conduit.

Although transparency (clarity of assessment purposes/formats) demonstrably enhances self-efficacy (Alkharusi, 2013), this study found no predictive effect on autonomous motivation. This paradox stems from dual practice failures. On the one hand, oversimplified tasks (e.g., 63% rote vocabulary drills) targeting struggling students nullify competence perception (McMillan & Workman, 1998), extinguishing deep learning drives. On the other hand, 78% classrooms withhold criteria beforehand (Tan et al., 2021), with teacher-dominated monolithic grading (e.g., reporting aggregate scores only) undermining transparency's discriminative role in efficacy beliefs (Alkharusi et al., 2014), severing the "transparency→efficacy→motivation" pathway.

e-ISBN: 978-629-98082-2-0

4.Conclusion

Anchored in Brookhart's (1997) classroom assessment environment theory, this study pioneers the examination of differential predictive effects of task perceptions on autonomous motivation in vocational EFL contexts using adapted scales. Structural equation modeling identified congruence, authenticity, and diversity as core motivational drivers, yielding three theoretical breakthroughs: validating the cross-contextual generalizability of Brookhart's tri-dimensional motivators (congruence-authenticity-diversity) in non-academic education tracks; revealing theoretical mediation fractures (efficacy-path disruption) for transparency/consultation in power-asymmetric classrooms, providing empirical leverage for contextual SDT recalibration; and demonstrating authenticity's motivational mediation through competence-need fulfillment (authenticity→competence→autonomous motivation), deepening mechanistic understanding of assessment-driven motivation. Consequently, pedagogical practice must harness authentic tasks to ignite intrinsic drive, ensure dynamic objective-assessment alignment, and design tiered tasks to navigate the structural challenge of polarized English proficiency in vocational cohorts.

The study has limitations. First, during the survey on task perception and autonomous learning motivation, students' inaccurate reports of their feelings may lead to data biases. Second, as a descriptive cross-sectional study, it can't establish causal relationships between assessment tasks, the assessment environment, and autonomous learning motivation. Future research should use experimental methods to validate these causal links. Third, the convenience sampling limited the subjects to students in Shanghai's secondary vocational English open classes, restricting the findings' generalizability to other regions. Future research should select more representative samples nationwide to confirm the conclusions.

References

Alkharusi, H. (2013). Canonical correlational models of students' perceptions of assessment tasks,2015 motivational orientations, and learning strategies. *International Journal of Instruction*, 6(1),21-38.

Alkharusi, H. A., & Al-Hosni, S. (2015). Perceptions of classroom assessment tasks: an interplay of gender, subject area, and grade level. *Cypriot Journal of Educational Sciences*, 10(1), 105-119.

Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2013a). The relationship between Students' Perceptions of the Classroom Assessment Tasks and academic achievement. *In 15th Annual International Conference on Education*, Athens, Greece.

Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2013b). The Impact of Students'

e-ISBN: 978-629-98082-2-0

Perceptions of Assessment Tasks on Self-efficacy and Perception of Task Value: A Path Analysis. *Social Behavior and Personality: an international journal*, 41(10), 1681-1692.

Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2014). Modeling the relationship between perceptions of assessment tasks and classroom assessment environment as a function of gender. *The Asia-Pacific Education Researcher*, 23, 93-104.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied measurement in education*, 10(2), 161-180.

Brookhart, S. M. (2014). *How to design questions and tasks to assess student thinking*. ASCD.

Chuane, Q., Shukor, S. S., Yuehong, T., & Xiaofen, Z. (2023). The relationship between motivation and English language test performance among secondary vocational schools' students in China. *Studies in English Language and Education*, 10(1), 280-302.

Browne, M. W. , & Cudeck, R. . (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230-258.

Deci, E. L., & Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4), 227-268.

Dorman, J. P., & Knightley, W. M. (2006). Development and validation of an instrument to assess secondary school students' perceptions of assessment tasks. *Educational Studies*, 32, 47-58.

Frey, B. B., Schmitt, V. L., & Allen, J. P. (2012). Defining authentic classroom instruction. Practical Assessment, *Research & Evaluation*, 17(2).

Gulikers, J. T., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2008). Authenticity is in the eye of the beholder: student and teacher perceptions of assessment authenticity. *Journal of Vocational Education and Training*, 60(4), 401-412.

Hao, T., & Pilz, M. (2021). Attractiveness of VET in China: A study on secondary vocational students and their parents. *Journal of Education and Work*, 34(4), 472-487.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Jin, Y., & Sun, H. (2020). Research on foreign language classroom assessment (2007-2018): Review and prospects. *Journal of Northeast Normal University (Philosophy and Social Sciences Edition)*, 5, 166-173.

Joe, H. K., Hiver, P., & Al-Hoorie, A. H. (2017). Classroom social climate, self-determined motivation, willingness to communicate, and achievement: A study of structural relationships in instructed second language settings. *Learning and individual differences*, 53, 133-144.

Li, Z., & Yang, Z. (2023). A study on the higher education choices and influencing factors of vocational school students: Based on a survey of 10,660 vocational school students

e-ISBN: 978-629-98082-2-0

nationwide. *Fudan Education Forum*, (01), 44-53.

Liu, X. (2024). Interactive teaching and its influence on academic motivation of students in a selected secondary vocational school in China. *Pacific International Journal*, 6(4), 85-91.

Lizzio, A., & Wilson, K. (2013). First-year students' appraisal of assessment tasks: implications for efficacy, engagement and performance. *Assessment & Evaluation in Higher Education*, 38(4), 389-406.

Macaskill, A., & Taylor, E. (2010). The development of a brief measure of learner autonomy in university students. *Studies in higher education*, *35*(3), 351-359.

McMillan, J. H., & Workman, D. J. (1998). Classroom Assessment and Grading Practices: A Review of the. *Psychology*, 84, 261-271.

Mouratidis, A., Michou, A., Sayil, M., & Altan, S. (2021). It is autonomous, not controlled motivation that counts: Linear and curvilinear relations of autonomous and controlled motivation to school grades. *Learning and Instruction*, 73, 101433.

Ng, B. L., Liu, W. C., & Wang, J. C. (2016). Student motivation and learning in mathematics and science: A cluster analysis. *International Journal of Science and Mathematics Education*, 14, 1359-1376.

Pajares, F., & Schunk, D. (2001). The development of academic self-efficacy. *Development of achievement motivation. United States*, 7, 1-27.

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. *Journal of Educational Psychology*, 103(3), 683.

Rabari, J. A. , & Indoshi, F. C. . (2011). Correlates of divergent thinking among secondary school physics students. *Educational Research*, 2, 982-996.

Ryan, R. M., & Deci, E.L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.

Tan, Y., Wang, Y., & Liu, X. (2021, June). Research on student assessment of secondary vocational schools in the Internet era. *In 2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)* (pp. 310-313). IEEE.

Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: toward a motivational model of high school dropout. *Journal of Personality and Social psychology*, *72*(5), 1161.

Van Dinther, M., Dochy, F., Segers, M., & Braeken, J. (2014). Student perceptions of assessment and student self-efficacy in competence-based education. *Educational Studies*, 40(3), 330-351.

Zhang, W., & Li, Y. (2019). A qualitative examination of classroom assessment in Chinese high schools from the perspective of self-regulated learning. *Frontiers of Education in China*, 14(3), 387-421.

e-ISBN: 978-629-98082-2-0

Zhang, W., & Li, Y. (2023). Development and validation of a questionnaire to assess classroom assessment from the self-regulated learning perspective. *Oxford Review of Education*, 49(6), 781-799.

Zimmerman, M. J. (2001). *The nature of intrinsic value*. Rowman & Littlefield Publishers.

e-ISBN: 978-629-98082-2-0

P092

Psychological Factors Influencing College English Learning: An Investigation of Learning Interest, Learning Motivation, and Interaction Anxiety among Chinese University Students

***Miao Yiran[1], Cheng Liying[2]And Yang He[3]**

[1] School of Foreign Languages, Chengdu University of Information Technology, China; School of Education, City University of Macau, Macau SAR
[2-3]School of Education, City University of Macau, Macau SAR
(E-mail: [1]m23092100840@cityu.edu.mo, [2]liyingcheng@cityu.edu.mo, [3]m23092100755@cityu.edu.mo)
**corresponding author: [1]m23092100840@cityu.edu.mo*

Abstract

In China, approximately 4 million undergraduate students are required to enroll in College English course yearly but a prevalent weakness in their English learning has been observed, which strongly correlates with students psychological factors. College English students are no longer compelled to study English for the purpose of the NCEE, resulting in the diminution of their pro-activity in English. Under this circumstance, English learning interest, learning motivation, and interaction anxiety have been documented to exert substantial influences on second-language acquisition (Mao, 2000; Hao et al., 2001; Guo, 2000). This study investigates intricate associations among learning interest, learning motivation, interaction anxiety, self-assessed English written level and English scores in NCEE. Data were collected from sophomore students at a public university in China using meticulously designed questionnaires. Preliminary analysis reveals moderate to high levels of learning interest (78%) and learning motivation (72.2%) among the valid responses, despite varying English proficiency levels. It's unexpected, given the observed lack of student engagement in actual class. Interaction anxiety was widespread, with 87% of students reporting varying degrees of anxiety. A strong positive correlation was found between self-assessed English written level and both learning interest ($r=0.446$, $p<0.01$) and learning motivation ($r=0.288$, $p<0.05$). Notably, no significant correlation was identified between NCEE scores and these 3 psychological factors, offering a new perspective for research that previous English learning experiences have limited influence on current learning psychological states, highlighting the potential to cultivate new attitudes towards English learning during college.

Keywords: College English students, learning interest, learning motivation, interaction anxiety

e-ISBN: 978-629-98082-2-0

1. Introduction

Numerous experts and scholars have empirically demonstrated that second language learners' learning motivation, interest, and anxiety all exert an influence on second language learning. According to past research, higher English learning motivation and interest can, to some extent, assist students in achieving better English learning outcomes. Meanwhile, moderate anxiety can enhance students' academic performance, but excessive anxiety can have a negative impact on learning. Based on previous literature, this study investigates intricate associations among learning interest, learning motivation, interaction anxiety, self-assessed English written level and English scores in NCEE, specifically examines the following hypotheses—Hypothesis 1: heightened learning interest and robust learning motivation exert significant positive effects on enhancing self-assessed English written level. Hypothesis 2: the prevalent teacher-centered instruction in College English classrooms induces interaction anxiety, thus, reduced interaction anxiety also positively improve self-assessed English written level. Hypothesis 3: high English scores in NCEE serve as a crucial factor in sustaining elevated learning interest and learning motivation, thereby providing a theoretical foundation for intervention strategies in educational practice.

2. Methods

This paper conducts a descriptive study to collect relevant data among College English students from sophomore and junior College English students at a university in Southwest China. Descriptive research does not test causal hypotheses; instead, it provides valuable information about who, what, when, where, and how related to a particular topic. Based on this information, researchers can propose hypotheses or determine the direction of future research. In this study, we describe (a) the interest in English learning among sophomore and junior College English students, (b) again, their motivation in English learning, (c) their interaction anxiety. Demographic questions encompass students' gender, age, place of origin, major, English learning experience and the self-assessment of their English written skill. The remainder of the survey is devoted to collecting data on students' motivation, interest in English learning, and interaction anxiety experienced after matriculation. All three scales exhibit good reliability and validity. In the collected questionnaires of the study, all data excepted the invalid ones were imported into SPSS software for analysis. The Cronbach's α coefficient of the learning interest scale reached 0.946, with a KMO value of 0.739; the Cronbach's α coefficient of the learning motivation scale reached 0.835, with a KMO value of 0.818; and the Cronbach's α coefficient of IAS reached 0.766, with a KMO value of 0.697. Therefore, all three scales exhibited good reliability and validity, demonstrating the objectivity and effectiveness of the questionnaires

e-ISBN: 978-629-98082-2-0

used in this study. The survey was distributed and managed through WenJuanXing, an online survey software widely used in China.

3. Results and Discussion

3.1 Students Demographics

3.1.1 College Entrance Examination English score summary

The survey was conducted among university English students from 8 different provinces and geographical origins and spanning 4 distinct majors. The findings reveal that 18.5% of the respondents failed to attain a score of 100 or higher in their College Entrance Examination English, while an overwhelming majority of 70% did not score 125 or above. Notably, 14.8% of the respondents managed to score over 130, resulting in an average score of 113.093 among the surveyed population as shown in Table 1.

Table 1: Table description

Basic indicators

	Minimum	Maximum	Mean□	SD□	Median□
College Entrance Examination English Score	81.000	139.000	113.093	14.352	113.000

Source: Compiled by researchers.

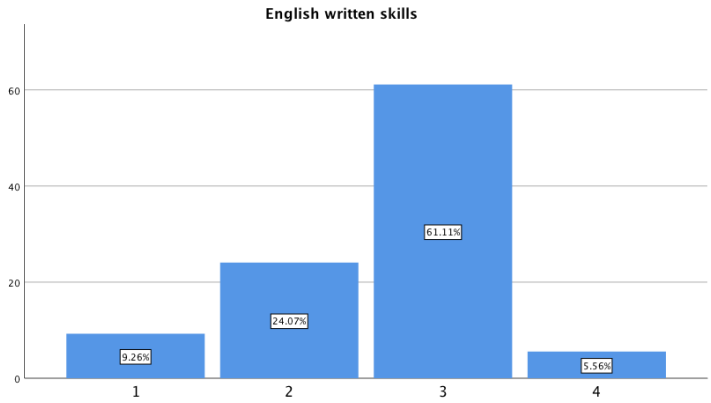
It is evidently apparent that among College English students, there exists a considerable disparity in their College Entrance Examination English Scores, with a substantial gap of 58 points between the highest and lowest performers. This significant variation poses a challenge to university English classroom instruction, as the prevailing question remains whether the currently utilized traditional teaching methodologies can adequately cater to the diverse learning needs of students across different proficiency levels.

3.1.2 Self-assessment of English written skill

Based on the data collected in this study, it is evident that in the self-assessment of English written skills spanning Levels 1 to 5, not a single student has rated themselves at Level 5. The majority of students perceive their English written abilities to be within the range of Levels 2-3. The specific data are illustrated in the figure provided as shown in Figure 1.

e-ISBN: 978-629-98082-2-0

Figure 1: Self-assessment of English written skill



Source: Compiled by researchers.

More than half of the respondents rated their English written skills at the third level, which signifies a moderate proficiency. This indicates that over 50% of the students consider themselves to have a foundational grasp of English writing but lack confidence, as only 5.56% of the students selected Level 4 or above. This observation is also attributed to the complexity inherent in the task of writing. Writing demands a considerable level of linguistic competence from students, hence, a certain degree of anxiety or reluctance towards this task is evident among them.

3.1.3 English learning interest, learning motivation and interaction anxiety

The data for the learning interest, learning motivation and interaction anxiety is illustrated in the Table 2：

Table 2: Table description

Basic indicators

	Minimum	Maximum	Mean□	SD□	Median□
Learning interest	49.000	167.000	105.185	18.292	107.000
Learning motivation	63.000	117.000	89.315	12.252	88.500
Interaction anxiety	26.000	60.000	42.630	7.693	43.000

Source: Compiled by researchers.

This study conducted normality tests on all three indicators, and the results are presented in the Table 3.

e-ISBN: 978-629-98082-2-0

Table 3: Table description

Normality Test Results

□	Mean□	SD□	Skewness	Kurtosis	Kolmogorov-Smirnov test	
					D-value in Statistics	p
Learning interest	105.185	18.292	0.137	2.524	0.094	0.280
Learning motivation	89.315	12.252	-0.172	-0.042	0.103	0.163
Interaction anxiety	42.630	7.693	0.012	-0.240	0.070	0.726

* $p<0.05$ ** $p<0.01$

Source: Compiled by researchers.

As evident from the table above, the sample size for the research data exceeds 50 in all cases, necessitating the application of the Kolmogorov-Smirnov (K-S) test. Specifically, for the indicators of learning interest, learning motivation, and interaction anxiety, none of them exhibited statistical significance ($p > 0.05$), suggesting the acceptance of the null hypothesis (which assumes normal distribution of the data). Consequently, the variables of learning interest, learning motivation, and interaction anxiety all exhibit characteristics of normality. From the data, we can ascertain that regarding learning interest, only 7% of college English students exhibit high interest (scores of 123.477 or above), while 14.8% demonstrate a low and negative interest (scores of 86.893 or below). The majority of students, accounting for 78%, maintain a moderate level of interest in English learning. As for learning motivation, merely 13% of college English students display high motivation (scores of 101.567 or above), whereas 14.8% also show a low and negative motivation (scores of 77.063 or below). Approximately 72.2% of students have moderate levels of motivation towards English learning. This indicates that currently, most non-English major students in this institution possess some degree of interest and motivation in learning English. If teachers can recognize this situation and implement measures to enhance students' interest and motivation in English learning, it may have a widespread positive impact. Regarding interaction anxiety, the total score ranges from 15 (indicating the lowest level of interaction anxiety) to 75 (indicating the highest level of interaction anxiety). Among the students surveyed in this study, 16.7% exhibited high levels of interaction anxiety, with scale scores ranging between 50 and 60. Conversely, 13% of students reported low levels of interaction anxiety. However, the majority of students, accounting for approximately 70.3% of the total survey respondents, experienced varying degrees of interaction anxiety, with scale scores falling between 35 and 50. Based on the overall data,

e-ISBN: 978-629-98082-2-0

approximately 87% of the students had interaction anxiety issues, which warrants attention and concern from institutions of higher education.

3.2 Correlation Analysis

Through a correlation analysis between the College Entrance Examination English scores and factors such as English learning interest, English learning motivation, and interaction anxiety, the following conclusions have been derived from Table 4:

Table 4: Table description

Pearson Correlation

		English Score for College Entrance Exam
Learning interest	Correlation Coefficient	0.170
	p-value	0.218
Learning motivation	Correlation Coefficient	0.159
	p-value	0.252
Interaction anxiety	Correlation Coefficient	-0.222
	p-value	0.107

* $p<0.05$ ** $p<0.01$

Source: Compiled by researchers.

Based on the results of the correlation analysis, the correlation coefficient between students' NCEE English scores and their interest in English learning is 0.170, which is close to 0, and the p-value is $0.218 > 0.05$, indicating that there is no correlation between NCEE English scores and interest in English learning. The correlation coefficient between students' NCEE English scores and their motivation for English learning is 0.159, also close to 0, with a p-value of $0.252 > 0.05$, suggesting no correlation between NCEE English scores and motivation for English learning. Furthermore, the correlation coefficient between students' NCEE English scores and their social anxiety is -0.222, nearing 0, and the p-value is $0.107 > 0.05$, thereby demonstrating no correlation between NCEE English scores and interaction anxiety. This result exhibits some differences from several previous research findings, where researchers often assumed a positive correlation between the performance in the College Entrance Examination English test and learning interest in, as well as learning motivation for, English learning. However, it is important to note that the respondents in this study are primarily sophomores, with some junior students included. These students have pursued non-English majors after entering university, which can also explain why their NCEE English scores from a year ago

e-ISBN: 978-629-98082-2-0

show no significant correlation with their current interest in and motivation for learning English, as well as their interaction anxiety. With these questions in mind, this study conducted a correlation analysis between students' self-assessment of their English written expression ability and their interest in learning, motivation for learning, as well as social anxiety. The analysis results are shown in the Table 5 below:

Table 5: Table description

Pearson Correlation

		self-assessed English written skills
Learning interest	Correlation Coefficient	0.446**
	p-value	0.001
Learning motivation	Correlation Coefficient	0.288*
	p-value	0.035
Interaction anxiety	Correlation Coefficient	-0.255
	p-value	0.063

* $p<0.05$ ** $p<0.01$

Source: Compiled by researchers.

Based on the analysis results, the correlation coefficient between students' self-assessed English written skills and their interest in learning is 0.446, showing significance at the 0.01 level, indicating a significant positive correlation between the two. The correlation coefficient between students' self-assessed English written skills and their motivation for learning is 0.288, exhibiting significance at the 0.05 level, suggesting a significant positive correlation between these two factors. The correlation coefficient between English written skills and interaction anxiety is -0.255, which is close to 0, and the p-value is $0.063 > 0.05$, indicating no correlation between English written skills and interaction anxiety. In other words, the better students' self-assessed English written skill, the higher their learning interest in and learning motivation for English, and vice versa. These results demonstrate that written expression ability may be one of the important factors influencing English learning. Based on the aforementioned results, the researchers also conducted a correlation analysis between interest in learning and motivation for learning. The final results indicate that the correlation coefficient between learning interest and learning motivation is 0.583, showing significance at the 0.01 level, thereby indicating a significant positive correlation between learning interest and learning motivation.

e-ISBN: 978-629-98082-2-0

3.3 Discussion

According to the statistical results, among the sophomores and juniors of this university taking college English courses, fewer than half of the students achieved scores above 120 (out of a total of 150) in the College Entrance Examination, with a gap of more than 50 points between the highest and lowest scores. After entering the university, these students pursue different majors, and the only systematic opportunity for them to learn English within the school is in College English classes. Therefore, employing the traditional teaching mode that primarily relies on lectures among student groups with such disparate English foundations is bound to greatly compromise teaching effectiveness. Especially in College English writing classes, writing is an output-oriented task that places certain demands on students' English proficiency. Although the process approach has been increasingly applied in college English writing instruction in China (Liang Yan, 2013), it remains teacher-centered, making it even more challenging for students with interaction anxiety to participate in classroom teaching. On the other hand, the descriptive data reveals over 85% of college English students hold a neutral to positive attitude towards English learning. This data surprised the researchers, because in most College English classrooms, student enthusiasm is not good. This also suggests that improving the teaching methods in college English classrooms may enhance students' classroom enthusiasm and, subsequently, teaching effectiveness. As a public and compulsory course, the teaching of College English should also take note of changes in learning situations, strive to meet the needs of the majority of learners, continuously update and improve teaching modes, and enhance teaching quality. College English teachers can explore ways to effectively enhance students' learning interest and motivation by trying out different teaching modes. When setting teaching objectives for college English, consideration can be given to incorporating the enhancement of students' learning interest and learning motivation into the assessment scope, which would contribute to their long-term development and lifelong learning. Lastly, according to the descriptive statistical results on interaction anxiety, 78.8% of students reported medium to high levels of anxiety based on their self-assessment scores. The scores indicate that a significant number of college students do experience varying degrees of interaction anxiety, which is consistent with the research findings of Jin Hua et al. (1986) and Peng Chunzi et al. (2004). Given this educational context, and considering the collaborative learning's inherent demands for interpersonal interaction and collaboration, this study further investigates the extent to which the educational methodologies and learning practices within the College English class may significantly mitigate or exacerbate their interaction anxiety.

4. Conclusion

This study primarily conducted descriptive statistical analysis alongside correlation analysis, compiled a table summarizing the collected data on College English students' interest in English

e-ISBN: 978-629-98082-2-0

learning, learning motivation, and current levels of interaction anxiety. The research indicates that the majority of college English students achieved above-average scores in the College Entrance Examination for English, yet there exists a significant disparity in their College Entrance Examination English scores, in other words, a substantial variation in their foundational English proficiency. Concurrently, even among those not majoring in English, most college students exhibit moderate to high levels of learning interest and learning motivation in English. Notably, this same group also experiences widespread interaction anxiety issues. These findings are consistent with numerous previous research outcomes and provide additional support, particularly in the context of interaction anxiety. Correlation analysis reveals that there is no clear positive correlation between College Entrance Examination English scores, self-assessed English written skills, and interaction. In other words, neither past English performance nor current self-assessed English written proficiency are correlated with college students' social anxiety. Therefore, improving English teaching methods in the classroom may serve as a means without pre-existing pressures to help alleviate the prevalent communication anxiety among college students.

References

Cai, Huiping, & Fang, Yan. (2006). Investigation and Analysis of the Current Situation of English Writing Teaching. *Foreign Languages and Their Teaching*, (09), 21-24.

Hao, Mei, & Hao, Ruoping. (2001). A Correlational Study of English Achievement, Achievement Motivation, and State Anxiety. *Foreign Language Teaching and Research*, (02), 111-115+160.

Herpratiwi, Herpratiwi, & Ahmad Tohir. (2022). Learning Interest and Discipline on Learning Motivation. International Journal of Education in Mathematics, *Science and Technology*, 10(2), 424–435. https://doi.org/10.46328/ijemst.2290

Jin, Yan, & Cheng, Liying. (2013). A Study on Psychological Factors Affecting the Validity of High-Stakes Tests. *Modern Foreign Languages*, (01), 62-69+109.

Lei, Huilin. (2012). A Brief Discussion on College English Writing Teaching Methods. *Science Education Guide (Mid-Month Edition)*, (06), 197-198.

Liang, Yan. (2013). *"Reading to Write": A Process Approach to College English Writing Teaching*. In China Informatization (Issue 10, pp. 293-293). College of Foreign Languages, *Hubei University of Science and Technology*, Xianning, Hubei, 437100.

Mao, Lingying. (2000). Current Situation and Countermeasures of College English Writing Teaching. *Journal of Chongqing University (Social Sciences Edition)*, (01), 98-100.

Ren, Fenglei. (2019). The Impact of Writing Motivation on College Students' English Writing Proficiency: The Mediating Effects of Individual and Environmental Factors. *Journal of Xiangnan University*, (01), 107-112.

e-ISBN: 978-629-98082-2-0

Tan, Xiaochun. (2010). A Study on the Interactive Effects of English Writing Motivation, Writing Strategies, and Writing Performance. *Journal of Hunan First Normal University*, (06), 61-66+109.

Wei, Chuxue. (2008). Constructivism Theory and Its Implications for College English Writing Teaching. *Higher Education Forum*, (03), 90-92+121.

Wei, Yan. (2010). A Review of College English Writing Teaching Research. *Journal of Hunan First Normal University*, (02), 71-74.

Yan, Lixia. (2012). Design and Analysis of English Learning Interest Scale for Vocational College Students. *Journal of Hubei Radio and Television University*, (05), 29-30.

Yang, Xiaoqiong, & Dai, Yuncai. (2015). A Practical Study of Autonomous Writing Teaching Mode Based on Pigai.org in College English. *Foreign Language Electronic Teaching*, (02), 17-23.

Zeng, Xihua, & Luo, Jiawen. (2012). A Study on the Characteristics of English Learning Motivation of Engineering College Students and Its Relationship with English Achievement. *Journal of Guangdong University of Technology (Social Sciences Edition),* (01), 72-76.

Zhang, Xuemei. (2006). An Investigation of the Current Situation of College English Writing Teaching. *Foreign Language World*, (05), 28-32.

e-ISBN: 978-629-98082-2-0

P099

CIPP Model-Based Evaluation and Optimization of Sino-Foreign Cooperative English Curriculum Systems

***Ning Song[1]; Jingyi Hu[2]**

School of Education, City University of Macau, China
(E-mail: [1]m24092100020@ cityu.edu.mo, [2]m24092100537@cityu.edu.mo)
**corresponding author: [1]m24092100020@ cityu.edu.mo*

Abstract

Employing the CIPP model, this inquiry investigates comprehensively and seeks enhancement of the College English curriculum within Sino-foreign cooperative education arrangements. It manifests an indispensable role these programs play in nurturing both linguistic adeptness and intercultural competencies among students. Through integrating mixed methodologies—encompassing surveys alongside interviews—the research illuminates the curriculum's efficiency in elevating English proficiency, harmonizing effectively with academic-professional purposes. Improvements remain necessary concerning discipline-specific content enrichment, augmented support for self-directed learning endeavors, plus equitable distribution of resources. Seen from the findings, imperative becomes continual curricular assessment to satisfy burgeoning demands collectively; practical suggestions emerge therein aimed at optimizing programmatic quality and securing graduates' readiness on a global scale.

Keywords: CIPP evaluation model; Sino-foreign cooperative education; college English curriculum enhancement; educational quality assessment

1. Introduction

Sino-foreign cooperative education has become integral to China's educational globalization, merging international resources with local contexts to develop students' professional and intercultural competencies (Gao, 2022; Miani & Picucci-Huang, 2023). The college English curriculum serves as a crucial conduit in these programs, requiring rigorous evaluation to ensure it meets evolving academic and career needs (Bao, 2023; Wu, 2024). This study applies the CIPP model (Stufflebeam, 1983) including encompassing Context, Input, Process, and Product dimensions to systematically assess curriculum alignment, resource allocation, instructional quality, and learning outcomes. Through mixed-methods analysis, we examine how effectively these English programs prepare students for global academic and professional environments, while identifying areas for targeted improvement. This study addresses four CIPP-based

e-ISBN: 978-629-98082-2-0

research questions: (1) curriculum objective alignment (Context); (2) resource adequacy (Input); (3) teaching effectiveness (Process); and (4) learning outcome relevance (Product) in Sino-foreign cooperative English programs.

2. Methods

This mixed-methods study evaluated English curricula in Sino-foreign programs using the CIPP model. A stratified random sample of 600 students yielded 218 valid questionnaires (36.3% response rate), representing diverse academic backgrounds. The 35-item instrument showed excellent reliability (α=0.992) and validity (KMO=0.970; 81.17% variance explained), validated by three TESOL experts. Semi-structured interviews (n=10, 15-20 minutes each) employed expert-reviewed protocols, with verbatim transcription. Methodological rigor was ensured through data triangulation (quantitative: descriptive statistics, factor analysis; qualitative: thematic analysis), member checking, and confidentiality protocols, enhancing validity while mitigating biases.

3. Results

This study employed a mixed-methods design combining quantitative questionnaire analysis (N=218; α=0.992) with qualitative thematic analysis (NVivo 12; κ=0.78) following Braun and Clarke's (2006) framework. The integration enabled methodological triangulation, with quantitative EFA revealing structural patterns and qualitative interviews providing contextual depth, collectively addressing all research objectives.

3.1 Research Question 1: Curriculum Objectives Alignment

Results showed strong agreement on curriculum clarity (89%, M=1.67) and program alignment (87.6%, M=1.71), though institutional integration scored lower (88%, M=1.74). Students endorsed IELTS 5.5 but reported discipline-specific vocabulary gaps, indicating needs for better vertical alignment and professional English integration.

Table 1: Participants' Feedback on the Clarity of English Curriculum Objectives and Context in Sino-Foreign Cooperative Education Programs.

No.	Item	Frequency and percentage	Agreement degree					N	M	SD
			Strongly agree	Agree	Neutral	Disagree	Strongly disagree			
1	The English curriculum has a clear and written vision and mission to guide the development of the Sino-foreign cooperative	Frequency	104	90	20	1	3	218	1.67	.776
		Percentage	47.7	41.3	9.2	0.5	1.4			

e-ISBN: 978-629-98082-2-0

	education program									
2	The English curriculum clearly defines its learning objectives to reflect the shared expectations of both partners	Frequency	97	94	23	1	3	218	1.71	.782
		Percentage	44.5	43.1	10.6	0.5	1.4			
3	The objectives of the English curriculum are consistent with the overall vision, mission, and educational goals of the Sino-foreign cooperative university	Frequency	95	91	28	1	3	218	1.74	.802
		Percentage	46.3	41.7	12.8	0.5	1.4			
4	The English curriculum sets specific learning objectives and plans learning paths based on the actual needs and background of learners (e.g., cultural background, language proficiency)	Frequency	95	88	32	1	2	218	1.74	.773
		Percentage	43.6	40.4	14.7	0.5	0.9			
5	The English curriculum considers the cognitive development stages of learners and sets corresponding learning objectives accordingly	Frequency	93	93	29	1	2	218	1.74	.773
		Percentage	42.7	42.7	13.3	0.5	0.9			
6	The English curriculum addresses the emotional needs of learners, such as motivation and interest, and reflects these in its objectives	Frequency	98	83	29	5	3	218	1.77	.865
		Percentage	45.0	38.1	13.3	2.3	1.4			
7	The English curriculum considers the physiological needs of learners (e.g., learning environment, time management) and reflects these in its objectives	Frequency	90	82	34	9	3	218	1.87	.919
		Percentage	41.3	37.6	15.6	4.1	1.4			
8	The English curriculum addresses the social needs of learners, such as cross-cultural communication skills and teamwork, and reflects these in its objectives	Frequency	94	89	31	2	2	218	1.76	.798
		Percentage	43.1	40.8	14.2	0.9	0.9			
9	The content and themes of the English curriculum aim to stimulate learners' curiosity and desire for exploration to adapt to the changing learning environment	Frequency	94	83	32	7	2	218	1.81	.869
		Percentage	43.1	38.1	14.7	3.2	0.9			
10	The English curriculum considers the latest developments in educational technology, such as online learning and blended learning, and reflects these in its	Frequency	90	95	27	3	3	218	1.78	.819
		Percentage	41.3	43.6	12.4	1.4	1.4			

e-ISBN: 978-629-98082-2-0

	objectives									
11	The English curriculum aims to build a knowledge system based on an international perspective to meet the learning needs in a globalized context	Frequency	92	95	27	2	2	218	1.75	.777
		Percentage	42.2	43.6	12.4	0.9	0.9			
		Total average						218	1.76	.814

3.2 Research Question 2: Resource Investment Adequacy

The evaluation showed positive resource investments (83.5%, M=1.79 for faculty; 85.8%, M=1.79 for facilities), though autonomous learning support was weaker (76.2%, M=1.86). Qualitative data revealed faculty strengths but gaps in practical materials, specialized labs, and equitable distribution, particularly for non-STEM disciplines.

Table 2: Participants' Feedback on the Relevance of English Curriculum Inputs in Sino-Foreign Cooperative Education Programs.

No.	Item	Frequency and percentage	Agreement degree					N	M	SD
			Strongly agree	Agree	Neutral	Disagree	Strongly disagree			
1	Satisfied with the resources and professional capabilities of the current English curriculum development team	Frequency	85	97	33	2	1	218	1.79	.761
		Percentage	39.0	44.5	15.1	0.9	0.5			
2	Observed or felt that English curriculum teachers frequently participate in professional training and development activities, such as teaching methods and techniques training, academic research and publication guidance, and domestic and international educational exchanges	Frequency	86	94	33	4	1	218	1.81	.791
		Percentage	39.4	43.1	15.1	1.8	0.5			
3	These professional training and development activities have helped improve teachers' teaching abilities	Frequency	85	102	29	1	1	218	1.77	.728
		Percentage	39.0	46.8	13.3	0.5	0.5			
4	The learning environment of the English curriculum is sufficient to stimulate interest and motivation for learning	Frequency	84	91	35	6	2	218	1.86	.849
		Percentage	38.5	41.7	16.1	2.8	0.9			
5	The English curriculum integrates multiple sources of knowledge, such as domestic and international	Frequency	90	102	24	1	1	218	1.72	.712
		Percentag	41.3	46.8	11.0	0.5	0.5			

e-ISBN: 978-629-98082-2-0

	textbooks, online resources, etc.	e								
6	The English curriculum ensures a safe physical learning environment, including classroom facilities and teaching equipment	Frequency	84	101	30	1	2	218	1.79	.763
		Percentage	38.5	46.3	13.8	0.5	0.9			
7	The English curriculum creates a safe, respectful, and supportive psychological learning environment	Frequency	87	102	26	0	3	218	1.76	.767
		Percentage	39.9	46.8	11.9	0	1.4			
8	The English curriculum refers to international advanced educational standards, such as those of relevant domestic and international educational associations	Frequency	82	98	33	3	2	218	1.83	.800
		Percentage	37.6	45.0	15.1	1.4	0.9			
9	The English curriculum keeps pace with the latest trends in global English education	Frequency	89	85	39	3	2	218	1.83	.835
		Percentage	40.8	39.0	17.9	1.4	0.9			
10	The English curriculum clearly defines the qualification requirements for foreign teachers	Frequency	89	96	28	3	2	218	1.78	.792
		Percentage	40.8	44.0	12.8	1.4	0.9			
11	Understand the specific requirements for foreign teachers' teaching qualifications, such as teaching experience, educational background, and professional certifications	Frequency	88	102	25	1	2	218	1.75	.747
		Percentage	40.4	46.8	11.5	0.5	0.9			
12	The English curriculum provides students with detailed and practical self-study guides or resources	Frequency	81	98	32	3	4	218	1.86	.849
		Percentage	37.2	45	14.7	1.4	1.8			
13	The English curriculum has established clear admission standards and assessment systems	Frequency	88	97	29	1	3	218	1.78	.796
		Percentage	40.4	44.5	13.3	0.5	1.4			
14	The English curriculum has set clear assessment standards and tracking mechanisms for student growth	Frequency	82	99	33	1	3	218	1.83	.801
		Percentage	37.6	45.4	15.1	0.5	1.4			
		Total average						218	1.80	.785

3.3 Research Question 3: Teaching Process Effectiveness

The evaluation demonstrated effective teaching processes, with 88.1% (M=1.74) endorsing learning support and 89% (M=1.71) valuing faculty feedback. While digital integration and

e-ISBN: 978-629-98082-2-0

career content were praised, challenges included limited independent learning effectiveness (13.3% moderate satisfaction) and low student involvement in curriculum design (3.2% dissatisfaction). Enhancing interactivity and participatory design could strengthen collaborative learning and student agency.

Table 3: Participants' Feedback on the Relevance of English Curriculum Processes in Sino-Foreign Cooperative Education Programs.

No.	Item	Frequency and percentage	Agreement degree					N	M	SD
			Strongly agree	Agree	Neutral	Disagree	Strongly disagree			
1	The English curriculum provides students with clear learning guidance and resource recommendations to support their English learning	Frequency	88	104	22	2	2	218	1.74	.749
		Percentage	40.4	47.7	10.1	0.9	0.9			
2	The English curriculum combines academic content with career development activities to promote students' overall development	Frequency	84	98	31	2	3	218	1.82	.811
		Percentage	38.5	45.0	14.2	0.9	1.4			
3	The English curriculum fully utilizes technological innovations (e.g., online learning platforms, smart tools), libraries, and language labs to meet students' learning needs	Frequency	84	90	34	6	4	218	1.88	.898
		Percentage	38.5	41.3	15.6	2.8	1.8			
4	The English curriculum promotes learning exchanges and cooperation among students, establishing a learning community	Frequency	86	97	31	0	4	218	1.80	.816
		Percentage	39.4	44.5	14.2	0	1.8			
5	In English classes, teachers frequently provide timely feedback to help students correct mistakes and improve learning	Frequency	93	101	21	0	3	218	1.71	.752
		Percentage	42.7	46.3	9.6	0	1.4			
6	The English curriculum has established effective communication mechanisms to regularly provide feedback on learning progress to students or parents	Frequency	87	86	37	4	4	218	1.86	.890
		Percentage	39.9	39.4	17.0	1.8	1.8			
7	English teachers provide ample assistance and support to students in and out of class to promote their learning	Frequency	91	100	24	0	3	218	1.73	.764
		Percentage	41.7	45.9	11.0	0	1.4			

e-ISBN: 978-629-98082-2-0

8	The English curriculum regularly assesses students' learning outcomes and provides personalized feedback and suggestions, adjusting teaching strategies accordingly	Frequency	88	97	28	2	3	218	1.78	.806
		Percentage	40.4	44.5	12.8	0.9	1.4			
9	The English curriculum activities include elements that encourage students to self-manage and self-control	Frequency	89	93	29	3	4	218	1.81	.853
		Percentage	40.8	42.7	13.3	1.4	1.8			
10	The English curriculum designs diverse teaching activities based on Gardner's multiple intelligences theory to meet different students' learning styles	Frequency	82	94	36	3	3	218	1.86	.839
		Percentage	37.6	43.1	16.5	1.4	1.4			
11	The English curriculum encourages students' independent learning and teamwork to promote their academic and career development	Frequency	87	96	29	3	3	218	1.80	.822
		Percentage	39.9	44.0	13.3	1.4	1.4			
12	Students or parents have the opportunity to participate in the teaching and evaluation process of the English curriculum, providing feedback and suggestions	Frequency	78	95	35	6	4	218	1.91	.889
		Percentage	35.8	46.3	16.1	2.8	1.8			
13	The English curriculum reasonably arranges study and rest time to ensure students' learning efficiency and physical and mental health	Frequency	83	87	38	5	5	218	1.91	.921
		Percentage	38.1	39.9	17.4	2.3	2.3			
14	The English curriculum allows students to participate in the formulation of course content and activity management to enhance their initiative and responsibility	Frequency	86	90	32	7	3	218	1.86	.881
		Percentage	39.4	41.3	14.7	3.2	1.4			
		Total average						218	1.82	.835

3.4 Research Question 4: Curriculum Outcomes Evaluation

The curriculum demonstrated strong outcomes across key domains: language proficiency (79.4%, M=1.83), career alignment (78%, M=1.90), and cross-cultural competence (84.9%, M=1.77). Students reported "dramatic improvements in speaking and writing" and valued the curriculum's professional relevance ("makes learning more meaningful"). Digital resources were deemed "invaluable," while practical applications received highest satisfaction (85.4%, M=1.92). Although autonomous learning (83.1%) and teamwork (84%) scored well, opportunities exist to enhance discipline-specific content and advanced language development. These findings confirm the curriculum's effectiveness while identifying targeted areas for

e-ISBN: 978-629-98082-2-0

refinement.

Table 4: Participants' Feedback on the Relevance of English Curriculum Outcomes in Sino-Foreign Cooperative Education Programs.

No.	Item	Frequency and Percentage	Agreement degree					N	M	SD
			Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree			
1	After completing the English curriculum, students feel that their English proficiency has significantly improved	Frequency	88	85	41	2	2	218	1.83	.828
		Percentage	40.4	39	18.8	0.9	0.9			
2	The content of the English curriculum is closely related to students' academic needs and future career development	Frequency	82	88	39	5	4	218	1.90	.898
		Percentage	37.6	40.4	17.9	2.3	1.8			
3	The English curriculum provides diverse learning resources (such as online platforms, library materials, etc.) to meet students' learning needs	Frequency	83	98	32	1	4	218	1.83	.828
		Percentage	38.1	45.0	14.7	0.5	1.8			
4	The English curriculum encourages students to engage in self-directed learning and self-management to enhance learning efficiency	Frequency	83	100	28	2	3	218	1.80	.770
		Percentage	38.1	45.9	14.7	0.5	0.9			
5	Teamwork activities in the English curriculum help improve students' social skills and collaborative abilities	Frequency	85	100	28	2	3	218	1.80	.801
		Percentage	39.0	45.9	12.8	0.9	1.4			
6	The English curriculum cultivates students' cross-cultural communication skills and international perspectives.	Frequency	92	91	30	3	2	218	1.77	.805
		Percentage	42.2	41.7	13.8	1.4	0.9			
7	The teaching methods in the English curriculum are flexible and varied, stimulating students' interest in learning	Frequency	88	98	27	3	2	218	1.78	.786
		Percentage	40.4	45.0	12.4	1.4	0.9			
8	The English curriculum offers ample practical opportunities for students to apply their knowledge to solve real-world problems	Frequency	81	87	40	6	4	218	1.92	.910
		Percentage	37.2	39.9	18.3	2.8	1.8			
9	The design of the English curriculum activities includes	Frequency	84	92	39	0	3	218	1.83	.815

e-ISBN: 978-629-98082-2-0

	elements that encourage students to self-manage and self-motivate, thereby enhancing their autonomous learning abilities	Percentage	38.5	42.2	17.9	0	1.4			
10	There is a high level of satisfaction with the English curriculum in the Sino-foreign cooperative education program	Frequency	79	98	35	3	3	218	1.87	.829
		Percentage	36.2	45.0	16.1	1.4	1.4			
		Total average						218	1.83	.827

4. Discussion

This CIPP-based evaluation identifies three key areas for curriculum enhancement in Sino-foreign English programs. First, while core language training proves effective (79.4% satisfaction), discipline-specific vocabulary gaps call for ESP integration, particularly for STEM/business majors. Second, despite strong infrastructure ratings (85.8%), equitable resource distribution remains crucial, especially for non-STEM students and specialized facilities. Third, high faculty support (89%) contrasts with moderate independent learning satisfaction (13.3%) suggesting project-based learning could boost engagement. The curriculum demonstrates particular strength in cross-cultural preparation (84.9%) and practical application (85.4%), with potential for further expansion through experiential learning opportunities. This study proposes a three-dimensional enhancement framework (layered instruction, equitable resources, participatory pedagogy) for optimizing transnational English curricula, with implications for maintaining global standards through localized CIPP evaluations.

5. Conclusion

This study's CIPP model evaluation yields three key contributions to transnational English pedagogy. First, it demonstrates how integrating language benchmarks (e.g., IELTS 5.5) with disciplinary requirements creates replicable academic-vocational frameworks. Second, it reveals how blended resources (e.g., CLIL modules with digital tools) enhance outcomes despite material constraints. Third, it establishes the value of dynamic CIPP-based evaluation cycles for maintaining curriculum relevance in initiatives like China's "New Liberal Arts." Methodologically, the innovative CIPP-guided mixed-methods approach effectively combines quantitative outcomes (79.4% satisfaction correlating with proficiency gains) with qualitative narratives. While offering practical solutions for addressing occupational instruction gaps and resource inequities, the study acknowledges limitations in its focus on short-term impacts within Sino-foreign programs. These findings provide valuable guidance for developing

e-ISBN: 978-629-98082-2-0

English curricula that support students' academic and global professional success.

Acknowledgement

We sincerely thank all participants, educators, and institutions for their invaluable contributions. Special gratitude goes to peer reviewers for their critical insights, and to our research team for their dedication. This study's success reflects our collective commitment to advancing Sino-foreign cooperative education.

References

Bao, Y. (2023). Research on teaching activities design of specialized English courses in Sino-foreign cooperative education based on 5C. *Advances in Education, 13*(03), 1056–1062. https://doi.org/10.12677/AE.2023.133167

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Gao, X. (2022). From internationalization to localization: A study on the development course of Sino-foreign cooperative education. *Advances in Education, 12*(04), 1363–1371. https://doi.org/10.12677/AE.2022.124211

Miani, M., & Picucci-Huang, S.-C. (Susan). (2023). Learning and Teaching in Transnational Education in China: Voices from Sino-Foreign Cooperative Universities. *Chinese Education & Society, 56*(5–6), 303–308. https://doi.org/10.1080/10611932.2024.2303912

Stufflebeam, D. L. (1983). The CIPP Model for Program Evaluation. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam, *Evaluation Models* (pp. 117–141). Springer Netherlands. https://doi.org/10.1007/978-94-009-6669-7_7

Wu, X. (2024). Research on learning strategies of bilingual courses under the background of Sino-foreign cooperative education. *Advances in Education, 14*(05), 224–229. https://doi.org/10.12677/ae.2024.145682

e-ISBN: 978-629-98082-2-0

P106

Navigating Topic Familiarity: Malaysian MUET Candidates' Challenges and Strategies in a High-Stakes Speaking Assessment

***Nurul Iman Ahmad Bukhari[1], Tengku Mohd Farid Tengku Abdul Aziz[2], Atirah Izzah Che Abas[3], Arifuddin Abdullah[4], Alla Baksh Mohamed Ayub Khan[5]**

[1, 3, 4]Faculty of Language Studies and Human Development, Universiti Malaysia Kelantan, MALAYSIA. [2]Akademi Pengajian Bahasa, Universiti Teknologi Mara Kelantan, MALAYSIA. [5]School of Languages, Literacies and Translation, Universiti Sains Malaysia, MALAYSIA.

(E-mail: [1]iman@umk.edu.my, [2]farid470@uitm.edu.my, atirah.ca@umk.edu.my, arifuddin@umk.edu.my,[5] allabaksh@usm.my)

**corresponding author: [2]farid470@uitm.edu.my*

Abstract

This qualitative study examines the critical influence of topic familiarity on speaking test performance among Malaysian University English Test (MUET) candidates, situating the analysis within broader language assessment challenges in the Global South. While topic familiarity is recognised as a key variable affecting test-taker outcomes globally, limited attention has been paid to the lived experiences and adaptive strategies of candidates in standardised high-stakes assessments across diverse sociolinguistic contexts. Addressing this gap, the research explores how topic familiarity interacts with cognitive, affective, and interactional challenges and identifies the coping mechanisms candidates employ to sustain coherence and fluency during examination. Twenty candidates aged 20–26, representing various Malaysian regions, participated in semi-structured interviews following their MUET speaking test sittings between 2022 and 2025. The interview protocol elicited detailed narratives concerning perceived topic familiarity, encountered difficulties in both individual and group speaking formats, and strategies adopted for unfamiliar or challenging prompts. Thematic analysis facilitated the identification of recurrent patterns and distinct experiences, offering a rich account of candidate agency within this assessment context. Findings reveal that topic familiarity promotes higher confidence, linguistic output, and engagement, while unfamiliar topics trigger heightened anxiety, cognitive blocks, and lexical retrieval difficulties. Candidates demonstrated considerable strategic resourcefulness, employing point structuring, advanced vocabulary rehearsal, repetition, personalised examples, and collaborative support in group settings. These adaptive strategies underscore the need for assessment practices attuned to local realities and for pedagogical interventions that address both linguistic proficiency and responsive test-taking skills. The study's implications are significant for language assessment

e-ISBN: 978-629-98082-2-0

policy, particularly in the Global South, where standardised English proficiency tests must reconcile equity, inclusivity, and contextual relevance. Recommendations include adopting equitable topic selection processes, fostering assessment literacy, and integrating targeted training on test preparedness and anxiety management. By foregrounding candidate perspectives and local context, this research contributes to debates on construct validity and fairness in high-stakes speaking assessments and calls for greater recognition of sociocultural factors shaping test performance in the Global South.

Keywords: topic familiarity; speaking assessment; MUET; language testing; high-stakes examinations

1. Introduction

The Malaysian University English Test (MUET) is a standardised English proficiency assessment required for admission to public universities across Malaysia. As part of its high-stakes examination framework, the speaking component requires candidates to deliver individual presentations and participate in group discussions, tasks that demand both confidence and communicative competence. Across global and, notably, Global South assessment contexts, topic familiarity has emerged as a crucial factor shaping test-taker outcomes, influencing not just language performance but also candidates' confidence, anxiety, and strategic behaviour.

Empirical evidence demonstrates that topic familiarity enhances fluency, lexical choice, and confidence in speaking assessments (Khabbazbashi, 2017; Abu Kassim & Zubairi, 2006). However, existing research seldom explores the lived experiences and adaptive strategies of candidates facing both familiar and unfamiliar topics. Recent work by Bukhari et al. (2023) specifically addresses this gap in the context of the Malaysian University English Test (MUET), finding that topic familiarity significantly influences performance and highlighting the necessity for formal teaching on commonly encountered topics to improve assessment validity. Bukhari et al.'s study, using many-facet Rasch measurement, also underscores the importance of designing tasks that account for differences in candidates' background knowledge to ensure fairness and reliability in high-stakes speaking assessments. Additionally, Mahmud and Najihah's research into the MUET demonstrates that the test has a substantial washback effect, motivating students to invest more effort in speaking skill development and intensifying their focus on communicative strategies during preparation. Significant concerns about construct validity arise when background knowledge—rather than language proficiency—potentially influences scores, raising questions of fairness and inclusivity (Weir, 2005; Bachman & Palmer,

e-ISBN: 978-629-98082-2-0

2010). Together, these studies reinforce calls for language assessment policies to prioritize equity and contextual relevance, particularly in diverse sociolinguistic environments.

This study addresses these gaps by examining how MUET candidates from diverse Malaysian regions experience topic familiarity during the speaking test, the nature of cognitive, affective, and interactional challenges they encounter, and the strategies they use to maintain coherence and fluency. By foregrounding these perspectives, the scope encompasses candidate responses to both test format and the realities of sociolinguistic diversity. Results offer significance for language assessment policy and pedagogical practice, emphasising the importance of equitable topic selection, increased assessment literacy, and support for strategic test preparation. The findings contribute to the ongoing debate on construct validity and fairness in high-stakes, multicultural speaking assessments.

Therefore, the research questions that guided this study are:

1. How does topic familiarity influence candidates' performance during the MUET speaking test?
2. What challenges do candidates experience during the MUET speaking test?
3. What strategies do candidates employ to overcome topic familiarity challenges during the MUET speaking test?

2. Methods

2.1 Participants

Twenty MUET candidates participated in this study, which included a diverse group of candidates who sat for the MUET speaking test in 2022-2025, with varying ages (20-26 years old) and from different regions in Malaysia, including Kelantan, Selangor, Negeri Sembilan, Kuala Lumpur, and Kedah.

2.2 Data Collection

Semi-structured interviews were conducted with all participants. The interviews focused on their experiences during the MUET speaking test, particularly regarding:

- The topics they received during the MUET Speaking test
- Perceptions regarding their topic familiarity in both Part 1 and Part 2 of the Speaking test.
- Challenges they faced during the test
- Strategies they employed to overcome these challenges

e-ISBN: 978-629-98082-2-0

2.3 Data Analysis

The interview transcripts were transcribed and subsequently analysed thematically, focusing on the three main research areas: topic familiarity, challenges, and strategies.

3. Results and Discussion

Interview analysis showed that MUET candidates faced a wide variety of speaking topics, including financial issues, social media, relationships, national interests, education, and health. Familiarity with the topic played an important role in shaping confidence and perceived difficulty. Candidates who had prior experience or preparation related to a topic felt more comfortable, while those facing unfamiliar topics reported increased nervousness and hesitation. Table 1 presents a summary of key findings from MUET speaking test candidates, organised by broad themes, sub-themes, observations, and illustrative evidence found.

Table 1: Summary of Key Findings on MUET Speaking Test Candidate Experiences

Theme	**Sub-theme**	**Observation**	**Illustrative Evidence**
Topic Familiarity	Topic Distribution	Candidates received topics on financial awareness, social media, relationships, helping others, national interests, education, food/culture, health/lifestyle.	"My topic was about part-time jobs, which I had done before."
	Perceived Topic Difficulty	Varied by familiarity: Topics described as easy, mixed, or requiring specialised knowledge.	"Relieved because it's not a strange topic, but still nervous about speaking."
	Influencing Factors	Prior preparation, real-life relevance, prior knowledge, personal interest affected perceived difficulty.	"I found it manageable because I often see this issue in my surroundings."
Cognitive Challenges	Mental blanks	Frequent loss of ideas or inability to continue speaking.	"My mouth wanted to speak, but my brain was thinking."
	Vocabulary limitations	Struggled to find wording, especially translating thoughts from Malay to English.	"Couldn't find words to translate, felt like wasting time."
	Difficulty elaborating	Repeated points due to inability to expand on ideas.	"Had moments of blankness and repeated the same point when speaking."

e-ISBN: 978-629-98082-2-0

Format-Related Issues	Time management	Insufficient time for presenting or group discussion.	"As candidate 1, didn't have much time to think or prepare."
	Group dynamics	Challenge in group work due to interruptions or conflicting ideas.	"When explaining, others would interrupt many times."
	Part preference	Varied: Some preferred individual task, others group discussion based on personal comfort and preparation.	"Group discussion was easier because of the main points provided."
Affective Challenges	Pre-test and performance anxiety	Anxiety, panic before or during the test; affected ability to write or speak fluent sentences.	"Panicked on seeing the question, had to repeat myself to avoid silence."
	Examiner-induced anxiety	Intimidation due to the examiner's fluency or style.	"Felt afraid that my delivery wouldn't match the lecturer's fluency."
Preparation Strategies	Structure and organisation	Use of frameworks (PEE, Introduction-Elaboration-Example-Conclusion, columns, brainstorming).	"Brainstormed and organised at least 3 ideas and a conclusion before speaking."
	Topic-specific preparation	Focusing on popular or past-year topics, preparing advanced vocabulary and sentences.	"Prepared bombastic words and power sentences ahead of the test."
Speaking Strategies	Anxiety management	Breathing techniques, self-talk, prayers, and smiling to mask nervousness.	"Took a deep breath before starting, tried not to show voice tremors."
	Recovery and repetition	Repeating or rephrasing points to maintain fluency during mental blocks.	"If blank, repeated previous point differently to remain involved."
	Collaborative support	Helped group members who struggled, aimed for discussion rather than debate.	"Tried to make it a discussion, not a debate, so everyone could contribute."
	Strategic point selection	Chose to discuss best-prepared points based on self-assessment and team dynamics.	"Discussed food vouchers first, saved education points for later when better prepared."
	Personalisation	Relating topic to own experience for more authentic, confident delivery.	"Related Part 2 topic to my own life to avoid mumbling or hesitation."

e-ISBN: 978-629-98082-2-0

Distinct cognitive challenges emerged in this study: candidates frequently described encountering mental blanks, difficulty elaborating points, and struggling to find appropriate words, particularly when translating from Malay to English. These findings echo and extend the results of earlier research that highlights topic familiarity as a facilitator of fluency, lexical choice, and confidence in speaking assessments (Khabbazbashi, 2017; Abu Kassim & Zubairi, 2006). However, while prior studies primarily quantified these effects, few delved into the specific lived experiences or coping methods candidates use when faced with unfamiliar topics. Our results bridge this gap by revealing that such cognitive blocks were often compounded by format-related issues—such as insufficient preparation time and complex group dynamics—which further impeded performance, a nuance often underemphasized in earlier literature.

Affective factors were central as well, with anxiety reported both before and during the test. This aligns with Bukhari et al. (2023), who found that topic familiarity significantly influences candidates' emotional states and advocated for explicit teaching of common topics to support student confidence and construct validity. Our study supports these claims but additionally documents the diversity of anxiety triggers—including examiner demeanour and delivery style—which were not foregrounded in previous work. The coping strategies identified here—such as structured use of frameworks, brainstorming, targeted vocabulary practice, and in-the-moment anxiety management—build upon the adaptive behaviours only superficially acknowledged in earlier studies. For instance, Mahmud and Najihah report that MUET exerts considerable washback, motivating focused preparation and the use of communicative strategies; our findings expand this by specifying the real-time tactics students deploy (e.g., repetition, peer support, personalisation) to sustain coherence under pressure.

Critically, our results reinforce concerns in the literature about construct validity and fairness, as raised by Weir (2005) and Bachman and Palmer (2010), with clear evidence that background knowledge can unduly affect scores. However, by foregrounding Malaysian candidates' voices, this study adds contextual depth to these debates, demonstrating how sociolinguistic realities and test format interact with emotional and cognitive demands to shape outcomes. Collectively, these findings underscore the need for assessment policies prioritizing equitable topic selection, explicit test strategy instruction, and robust anxiety management, echoing and elaborating on the emerging consensus in recent scholarship for a more inclusive and contextually attuned approach to language assessment.

e-ISBN: 978-629-98082-2-0

4. Conclusion

This study enhances understanding of Malaysian MUET candidates' experiences by highlighting the central role of topic familiarity in speaking test performance. The results show that candidates' confidence and language output are positively shaped when engaging with familiar topics, while unfamiliar subject matter introduces distinct cognitive, format-related, and affective challenges. Despite these difficulties, candidates demonstrate resourcefulness by employing structured preparation, advanced vocabulary, repetition, personal examples, and collaborative strategies to maintain fluency and coherence.

The connection between topic familiarity and test-taker experiences underscores important implications for language assessment practice in the Global South. Findings suggest the need for equitable topic selection, increased assessment literacy, and targeted training on anxiety management and test preparedness. By foregrounding lived experiences and adaptive strategies, this research advocates for fairer, more responsive speaking assessment policies and calls for greater consideration of sociocultural context in both test design and pedagogy.

References

Abu Kassim, N. L., & Zubairi, A. M. (2006). Interaction between test-taker characteristics, task facets and L2 oral proficiency test performance. *Educational Research*.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford University Press.

Bukhari, N. I. A., Ismail, L., Abu Kassim, N. L., Razali, A. B., Noordin, N., & Mohd Noh, M. F. (2023). Topic familiarity effects on performance in speaking assessment tasks. *Arab World English Journal, 14*(4), 213-232.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*.

Mahmud, N., & Najihah, M. (2018). Investigating the washback effect of the MUET as a university entry test on students in Malaysia. *International Journal of Learning, Teaching and Educational Research*.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Palgrave Macmillan.

e-ISBN: 978-629-98082-2-0

R002

The Application of Generative AI in Formative Assessment for English Writing Skills in K-12 Education: Challenges and Opportunities

Zihan Sun

City University of Macau

Abstract

Generative artificial intelligence (AI) is reshaping formative assessment in language education by offering personalised, near-real-time feedback at classroom scale. This pilot study compares three engines—Feixiang AI Composition, iFLYTEK Spark Writing, and GrammarlyGO—across four VEPE dimensions (Validity/Effectiveness, Equity, Pedagogy, Ethics). A corpus of 300 English essays produced by 300 Grade-7–9 students in three mainland-China secondary schools was double-scored by AI and six trained teachers. AI feedback shortened marking time by 29 % and raised learner-engagement scores ($\Delta = 0.52$, $p < .01$). Yet discourse-level agreement remained moderate ($r = .62$), and Rasch DIF ($\Delta\beta \approx 0.38$ logits) indicated residual bias against lower-proficiency writers. Privacy audits confirmed PIPL compliance for the two Chinese systems, whereas Grammarly retains user text on overseas servers. We propose a layered-intervention model: AI for micro-level diagnosis, teachers for macro-level coherence and affective support. The study highlights both the promise and the limits of AI-mediated feedback, offering practical guidance for educators and tool designers and mapping future research priorities.

Keywords: Generative AI; Formative Writing Assessment; EFL Learners; Automated Feedback; Validity and Equity Framework

Extended Abstract

Purpose

Generative large-language-model (LLM) systems have begun to reshape formative writing assessment, yet empirical evidence on their classroom impact—particularly in K-12 Chinese contexts—remains sparse. This mixed-methods study evaluates three engines—**Feixiang AI Composition**, **iFLYTEK Spark Writing**, and **GrammarlyGO**—through the four-dimension **VEPE** lens (Validity/Effectiveness, Equity, Pedagogy, Ethics). Two research questions guide the work:

e-ISBN: 978-629-98082-2-0

1. What VEPE-framed opportunities and challenges emerge in the 2022-2025 literature on AI-mediated writing feedback?
2. How do the three tools perform on VEPE dimensions in real Shenzhen classrooms, and how do teachers perceive them?

Methods

A PRISMA-ScR review (Jan 2022–Apr 2025) identified **22 empirical studies**. The field pilot involved **300 Grade 7-9 learners** and **six English teachers** across **three Shenzhen secondary schools**. Each tool scored 100 expository essays; teachers double-scored a 20-essay anchor set ($\kappa = .83$), completed a 6-item NASA-TLX survey ($\alpha = .87$), and participated in semi-structured interviews. Quantitative indices—human–AI agreement, Hedge's *g*, Rasch DIF, workload logs—were computed in R 4.4.0 with 95 % CIs; interview and feedback-episode transcripts (n = 75) were double-coded ($\kappa = .84$) and deductively mapped to VEPE categories.

Results

Turnaround time averaged 20 s (Spark), 60 s (Feixiang), and 45 s (Grammarly), versus a 48-h teacher baseline.

Engagement rose significantly in AI classes ($\Delta = 0.52$, $p < .01$).

Validity: overall r = .84, $\kappa = .78$; coherence r = .62 indicates a discourse "blind spot."

Equity: DIF $\Delta\beta \approx 0.34–0.40$ logits favoured higher-proficiency writers.

Pedagogy: marking-time fell 29 %; yet only 33 % of teachers would "always" accept AI suggestions on higher-order traits.

Ethics: Feixiang and Spark met PIPL data-governance rules; Grammarly defaulted to overseas storage.

Conclusions

No single engine is best-in-class. Feixiang scored highest on pedagogy and ethics; Spark excelled in holistic validity but lagged on equity; Grammarly provided balanced mid-tier performance with privacy caveats. A **three-layer intervention model** is proposed: AI for micro-level diagnosis, teachers for macro-level discourse mentoring, learners for reflective consolidation.

e-ISBN: 978-629-98082-2-0

Implications

Open, curriculum-aligned "GLUE-Write-CN" corpora and mandatory *pedagogical explainability statements* could address validity and bias gaps. A 15-hour practice-embedded PD sequence is recommended to build teacher trust and orchestration capacity. Future studies should test VEPE metrics in rural or cross-national settings and across narrative or multimodal genres to gauge longitudinal effects on learner self-regulation.

e-ISBN: 978-629-98082-2-0

R006

The Influence of AI-Driven Writing Assistants on Students' Attitudes Towards Writing Skills and Academic Honesty

Nur Ain Amani Mohd Azmi[1]

[1]Akademi Pengajian Bahasa, Universiti Teknologi MARA Cawangan Selangor, MALAYSIA.
(E-mail: [1]2023232734 @student.uitm.edu.my)
**corresponding author: [1]2023232734 @student.uitm.edu.my*

Abstract

The growing use of AI-driven writing assistants in academic settings raises concerns about their impact on students' writing proficiency and academic honesty. The reliance on AI tools cannot be addressed simply by integrating them into academic writing without clear guidelines. There is an urgent need to understand how students use these tools, their perceived benefits and drawbacks, and the factors influencing their adoption. Thus, a strategic approach is necessary to balance AI's role in enhancing writing skills while maintaining academic integrity. This research aims to examine students' attitudes towards AI-driven writing assistants, focusing on their influence on writing proficiency, academic honesty, and the key factors influencing their usage. The study specifically investigates undergraduate students at UiTM Shah Alam, applying the Technology Acceptance Model (TAM) to assess their attitudes. A mixed-methods approach was employed, incorporating two focus group discussions and survey data from 244 students. The findings indicate that students primarily use AI tools for grammar correction, sentence restructuring, and time management, but do not strongly associate them with long-term writing improvement. Additionally, while AI is seen as useful for plagiarism checks, concerns remain about ethical misuse and overreliance. The study suggests that clear academic policies and AI literacy programmes are needed to promote responsible AI use. The findings contribute to understanding AI's role in academic writing and provide recommendations for integrating AI tools effectively without compromising students' writing proficiency and academic honesty.

Keywords: Artificial Intelligence in education; AI-driven writing assistants; writing proficiency; academic dishonesty; Malaysia tertiary education

1. Introduction

Artificial Intelligence (AI) is advancing rapidly and changing many industries, including education. As AI grows, schools and universities have started using AI-driven writing tools like

e-ISBN: 978-629-98082-2-0

Grammarly, Quillbot, and GPT-based platforms to assist in writing. These tools do everything from fixing grammar and spelling to content generation, shifting the teaching and learning discourse for students and teachers. Current education reforms focusing on educational resources, gamification, and personalised learning provide abundant opportunities for the development of educational AI applications (Zhai et al., 2021). Malaysia's educational landscape has also embraced the integration of AI, planning to build several national technological institutes, including various AI facilities dedicated to education, research, and development of the field (Chatterjee & Bhattacharjee, 2020). Narrowing the implications of AI tools on writing skills, it is still a subjective and abstract area whether the use of AI aids or hinders students' writing proficiency. Students have long been relying on forms of AI, from the red or blue squiggly lines in word processors to the more progressive software, suggesting that AI-driven writing assistants are just another integral part of one's writing (Escalante et al., 2023). Hence, from another perspective, the consolidation of AI-driven writing tools and academic writing produces contrasting attitudes and outcomes. Closely related to the topic is the issue of academic honesty, which involves maintaining transparency, integrity, and ethics in academic work. The emergence of sophisticated AI models poses future problems with distinguishing between human-written and AI-generated text, necessitating the development of more advanced detection tools (Grassini, 2023). A central ethical issue in academia involves the misuse of AI tools, particularly concerning how extensively the tools are used.

A major concern is that relying too much on AI could weaken students' ability to write independently. When students accept AI-generated suggestions without thinking critically, they may struggle to develop their arguments and improve their writing skills (Halaweh, 2023). Most writing research on AI focuses on theory rather than practical effects in higher education. While studies on AI reliance mostly look at decision-making (Klingbeil et al., 2024), little is known about how much students depend on AI writing tools for academic work. The long-term effects of such tools on students' ability to write and edit independently are also yet to be comprehended. Without clear guidance on balancing AI use with traditional writing skills, students risk becoming too dependent on AI and losing opportunities to practise critical thinking. As AI-driven writing assistants become more common, concerns about originality, authorship, and academic honesty are growing (Miao et al., 2023). Students often struggle to distinguish between acceptable AI support and unethical writing practices, especially since there is a lack of clear guidelines for AI-supported writing (Gustilo et al., 2024). The readily available AI technology raises plagiarism concerns because students can present work generated by AI without acknowledging the source, thus making it hard to distinguish between acceptable academic assistance and what crosses the line into academic misconduct.

e-ISBN: 978-629-98082-2-0

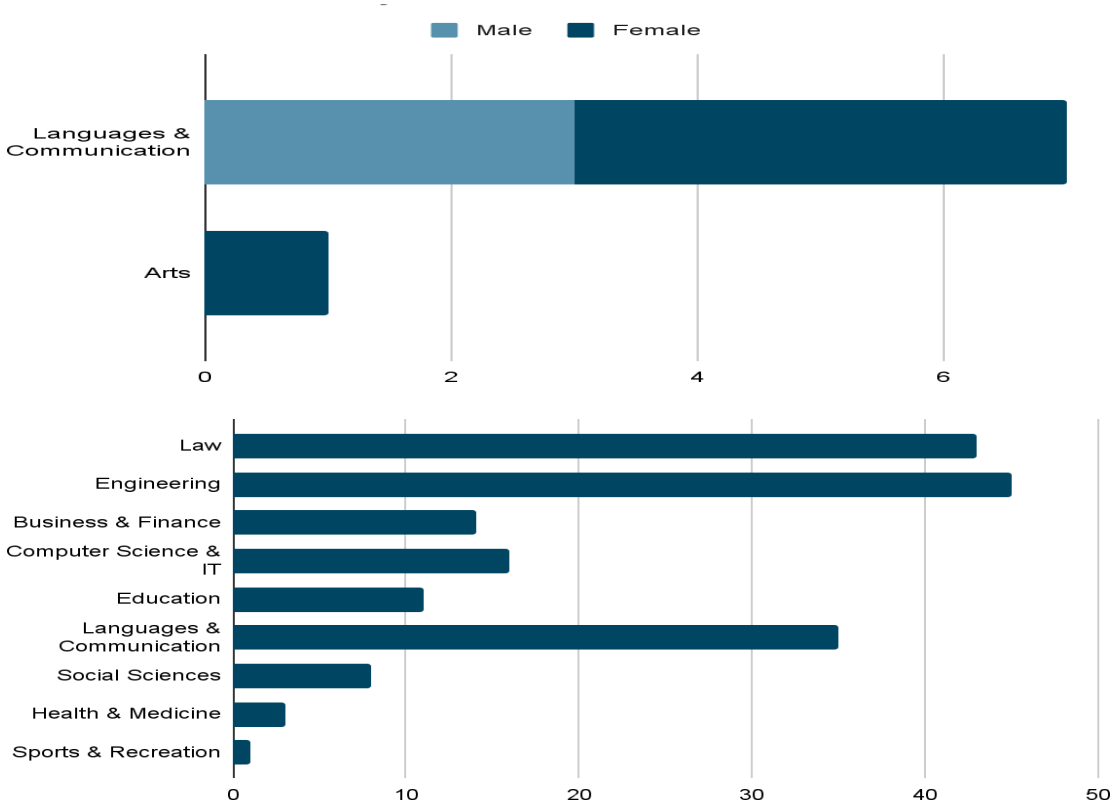
Despite increasing global interest in AI in education, Malaysian research remains limited, focusing mainly on students' acceptance rather than its impact on writing proficiency and academic honesty. As international studies grow, there is a pressing need for local research to examine whether AI supports skill development or leads to overreliance and dishonesty. This study aims to address this gap by exploring Malaysian students' attitudes towards writing proficiency and academic integrity in the context of AI use. Therefore, the purpose of the study is to explore: (1) students' attitudes towards AI-driven writing assistants on writing proficiency, (2) students' attitudes towards AI-driven writing assistants on academic honesty, (3) relationship between students' attitudes towards AI- driven writing assistants on writing proficiency and academic honesty, and (4) factors influencing use of AI-driven writing assistants.

2. Research Methodology

This study employed a mixed-methods design with an exploratory sequential approach to explore UiTM undergraduate students' attitudes towards AI-driven writing assistants concerning writing proficiency and academic honesty, while also examining the factors influencing these attitudes. The design and approach were chosen as they provided a comprehensive understanding of the research objectives by leveraging the strengths of both qualitative and quantitative methodologies. Qualitative data were first collected through semi-structured interviews to better comprehend students' views and experience of AI writing assistants, focusing on writing competency and views on academic integrity. The findings were later utilised to guide the construction of a quantitative survey to determine students' attitudes on a wider scale. This step-by-step approach allowed for a deeper understanding of key themes and ensured a thorough analysis of the factors influencing students' perspectives on AI-driven writing assistants.

e-ISBN: 978-629-98082-2-0

Figure 1:



The semi-structured interview was conducted on eight participants. Focus group discussions were conducted due to the interactive nature that encouraged participants to discuss and debate their opinions comfortably. Figure 3.1 shows participants' demographics by gender and academic background. Languages & Communication had the highest representation (7 participants: 3 males, 4 females), while the Arts field had only one female participant. Meanwhile, a minimum of 382 respondents were targeted for the quantitative data, based on a 95% confidence level and an 8% margin of error using the Raosoft calculator. Simple random sampling was employed to give each member of the population an equal chance of selection, reducing biases such as sampling and selection bias (Thomas, 2020). Figure 3.2 presents the distribution of survey respondents by field of study. The highest number of participants was from Engineering (45), Law (43), and Languages & Communication (35), and other fields with fewer respondents.

For the qualitative phase, two focus group interviews were conducted using a semi-structured format, with questions adapted from the Technology Acceptance Model (TAM). For the quantitative phase, a survey was designed based on TAM and insights from the interviews. It measured students' perceived ease of use, usefulness, and attitudes toward AI tools, using a 4-point Likert scale to avoid neutral responses. A pilot study with 31 participants confirmed the

e-ISBN: 978-629-98082-2-0

instrument's reliability, with a Cronbach's Alpha of 0.910, indicating strong internal consistency. The data collection for the qualitative phase took place on an online platform and was recorded with participant consent, with two focus group discussions, each with four participants, using a discussion guide focused on key themes related to AI writing assistants, writing proficiency, and academic honesty. For the quantitative phase, 244 responses were collected from UiTM Shah Alam students via email and WhatsApp, ensuring a wide reach through familiar platforms. The online questionnaire was available for six months, after which data collection ended, and analysis commenced.

Qualitative data from focus groups were transcribed and analysed using thematic analysis via NVivo14, identifying key themes on AI writing tools, writing proficiency, and academic honesty. These insights then informed the design of the quantitative survey. Quantitative data were analysed using SPSS (Version 29), applying descriptive statistics and Pearson correlation to examine the relationship between students' attitudes towards AI tools, writing skills, and academic honesty.

Table 1: Research Objectives and Data Analysis on AI-Driven Writing Assistants

Research Objective(s)	Data analysis
To determine UiTM undergraduate students' attitudes towards AI-driven writing assistants on writing proficiency.	Descriptive Statistics
To determine UiTM undergraduate students' attitudes towards AI-driven writing assistants on academic honesty.	Descriptive Statistics
To examine the relationship between UiTM undergraduate students' attitudes towards AI-driven writing assistants on writing proficiency and academic honesty.	Pearson Correlation Analysis
To examine factors that influence UiTM undergraduate students' use of AI-driven writing assistants.	Thematic Analysis

3. Results and Discussion

3.1 Attitudes towards AI-driven writing assistants on writing proficiency

Students primarily use AI writing assistants for grammar correction and sentence restructuring (A12, M=3.47) and perceive them as time-saving tools (A5, M=3.41; A13, M=3.38; A1, M=3.38). They also find them generally usable and accessible (A14, M=3.35; A11, M=3.29). However, AI tools are not strongly associated with significant writing improvement or productivity (A2, M=3.19; A3, M=3.13). The results indicate that students perceive AI writing assistants as useful for basic linguistic support, particularly in grammar correction, sentence restructuring, and saving time. However, they do not strongly associate these tools with major improvements in writing proficiency or productivity.

e-ISBN: 978-629-98082-2-0

3.2 Attitudes towards AI-driven writing assistants on academic honesty

Students view AI-driven writing assistants as moderately useful for plagiarism detection (A17, M=2.97) and ensuring originality (A18, M=2.95; A16, M=2.92) but have lower confidence in their role for citation accuracy (A15, M=2.73) and overall academic honesty (A8, M=2.50). While AI tools are acknowledged for supporting integrity under pressure (A9, M=2.71; A10, M=2.71), their effectiveness in enhancing credibility (A7, M=2.45) and reducing plagiarism risks (A6, M=2.32) is perceived as limited. The results suggest that while students find AI writing tools somewhat helpful for checking plagiarism and ensuring originality, their overall confidence in AI's role in maintaining academic honesty remains moderate to low.

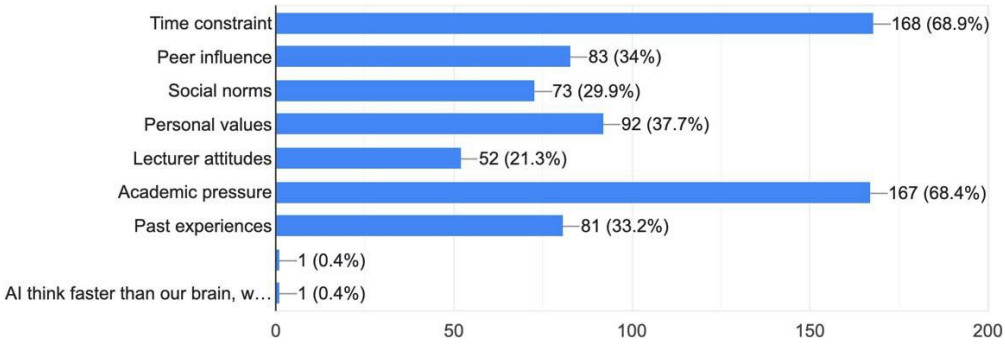
3.3 Relationship between attitudes towards AI-driven writing assistants on writing proficiency and academic honesty

The study explored the relationship between ATTITUDE_WP (Attitudes towards Writing Proficiency) and ATTITUDE_AH (Attitudes towards Academic Honesty). The study demonstrates a moderate, positive relationship between the two factors with a Pearson correlation of 0.386, significant on the 0.001 level ($p < 0.001$). A moderate, positive relationship indicates that when students feel more positively towards their writing abilities, they feel more positively towards being honest in class. This suggests that the relationship observed in this study is likely to exist in the broader student population as well. Next, for this relationship, ATTITUDE_WP is the independent variable, as it reflects students' perceptions of writing proficiency and its role in their academic experience. ATTITUDE_AH is the dependent variable, as it measures how students' attitudes toward writing skills potentially influence their views on academic honesty.

3.4 Factors Influencing Use of AI-driven Writing Assistants

The analysis revealed several factors that shape the way undergraduate students view AI writing assistants: (1) time constraints, (2) peer influence, (3) personal values, (4) lecturer attitudes, (5) academic performance, and (6) past experiences.

Figure 1: Factors influencing use of AI-driven Writing Assistants



e-ISBN: 978-629-98082-2-0

Figure 6.1 illustrates the factors influencing students' attitudes towards academic honesty in their academic work. The most significant factors include time constraints (68.9%) and academic pressure (68.4%), indicating that students may resort to AI writing tools due to workload and deadlines. Other notable influences include personal values (37.7%), peer influence (34%), and past experiences (33.2%). Social norms (29.9%) and lecturer attitudes (21.3%) also play a role, though to a lesser extent. A tiny percentage (0.4%) indicated other reasons.

Time constraints and academic pressure were the most cited factors, with 68.9% and 68.4% of respondents, respectively, indicating that these contributed to their AI use. Students often turned to AI tools during peak academic weeks or when juggling multiple assignments, seeing them as a means to save time and reduce stress. Robinette et al. (2016) similarly found that time constraints increase users' dependence on AI and affect how they verify content. Ethical concerns arise, particularly around academic honesty, though some students justify AI use due to deadline urgency. For English language learners, time constraints limit writing practice, making AI's time-saving benefits appealing (Song & Song, 2023). This highlights a tension between practical use and ethical boundaries in academic contexts.

Peer influence plays a significant role in shaping students' use of AI tools. Many students reported discovering and adopting AI through friends, leading to its normalised use in group settings. Some described cross-checking AI-assisted work with peers, reflecting a collaborative environment where AI use is expected. Others began using AI based on peer recommendations, showing how suggestions influence tool adoption. This supports Clasen and Brown's (1985, as cited in Xu et al., 2023) concept of peer pressure influencing behaviour.

Attitudes toward AI in writing differ based on personal values. Some view it as a threat to originality and critical thinking, echoing concerns by Marzuki et al. (2023) about overreliance weakening these skills. Others see AI as a supportive tool that enhances writing without compromising integrity, in line with Halaweh (2023) and Al-Abdullatif & Alsubaie (2024), who highlight its potential as a valuable educational resource. These contrasting views reflect how individual beliefs shape students' acceptance and use of AI in academic settings.

However, lecturer attitudes appeared to have a lesser impact; At the same time, some students mentioned that their lecturers discouraged overreliance on AI, and institutional policies on AI use were often unclear or inconsistently communicated. Lecturers who promote academic integrity and critical thinking while supporting AI use can help students avoid issues like plagiarism or over-dependence on technology. When AI tools are perceived as threats to academic integrity, students may misuse them in an attempt to bypass academic expectations.

e-ISBN: 978-629-98082-2-0

Just as Chan & Tsi (2023) inform, generally, AI can influence teachers' relationships, trust, and communication by disrupting their interactions with students, colleagues, and parents. Nevertheless, the findings provide evidence supporting existing literature that highlights lecturer attitudes as a key influence on students' use of AI tools in academic work.

Next, the findings suggest that academic performance significantly influences students' use of AI-driven writing assistants as they seek to meet academic expectations and enhance their work quality under pressure. Academic performance pressures drive students to use AI writing tools for organising ideas, refining writing, and managing deadlines, especially during writer's block (Washington, 2023; Rahman et al., 2022). Despite these advantages, overreliance on AI may hinder independent writing skills and critical thinking abilities. AI reliance can reduce engagement in higher-order cognitive processes and limit opportunities for students to construct well-developed arguments independently (Klingbeil et al., 2024). It shows that while AI writing tools can support students' academic performance, it is important to balance AI assistance with independent learning to develop long-term skills. Future research should focus on how to use AI in academic writing effectively without weakening critical thinking, originality, or deeper cognitive development.

Finally, prior interactions, whether positive or negative, influence students' perceptions, confidence, and willingness to engage with AI in academic work. Negative experiences often lead to caution and skepticism. Some students feel that excessive use of AI tools has lessened their self-confidence, so they are apprehensive about relying on them for academic assignments. This feeling of intellectual decline highlights the balance between AI's convenience and the possible weakening of cognitive skills, including problem-solving and critical thinking, as discussed by Burkhard (2022). In contrast, positive experiences foster more favourable attitudes. These findings support existing literature that highlights experience as one of the key factors influencing AI use.

4. Conclusion

This study found that UiTM undergraduate students generally view AI-driven writing assistants as beneficial for improving writing efficiency, though their attitudes toward academic honesty remain cautious. Guided by the Technology Acceptance Model (TAM), the findings highlight that perceived usefulness and ease of use significantly influence students' engagement with AI tools. However, concerns over overreliance on AI raise ethical and pedagogical implications, particularly regarding reduced creativity and critical thinking. Educators are encouraged to embed AI literacy into coursework to promote responsible and reflective use. Clear institutional policies are needed to guide students in using AI as a learning support rather than a shortcut.

e-ISBN: 978-629-98082-2-0

While the study offers valuable insights, its findings are limited by sample size and scope, suggesting that broader research is necessary to fully understand AI's impact in varied educational contexts.

Acknowledgement

First and foremost, I am deeply grateful to Allah, the Almighty, the Most Gracious, and the Most Merciful, for His endless blessings and guidance throughout this journey. May peace and blessings be upon Prophet Muhammad (peace be upon him), his family, and his companions.
I would like to express my heartfelt appreciation to my supervisor, Dr. Norhaslinda Hassan, for her unwavering support, invaluable guidance, and patience from the beginning of this research. Her insightful advice and encouragement have been a pillar in shaping this thesis. Above all, I dedicate this thesis to my beloved parents, Mohd Azmi Omar and Jamilah Mohamad, whose unwavering love, sacrifices, and encouragement have been my greatest source of strength.
May this thesis be of benefit to future scholars and researchers.

References

Al-Abdullatif, A. M., & Alsubaie, M. A. (2024). ChatGPT in learning: Assessing students' use intentions through the lens of perceived value and the influence of AI literacy. *Behavioral Sciences, 14*(9), 845. https://doi.org/10.3390/bs14090845

Burkhard, M. (2022). Student perceptions of AI-powered writing tools: Towards individualized teaching strategies. In D. G. Sampson, D. Ifenthaler, & P. Isaias (Eds.), *Proceedings of the 19th International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA 2022)* (pp. 72–81). IADIS. https://doi.org/10.33965/celda2022_202207l010

Chan, C. K. Y., & Tsi, L. H. (2023*). The AI revolution in education: Will AI replace or assist teachers in higher education?* arXiv. https://arxiv.org/abs/2305.0118

Chatterjee, S., & Bhattacharjee, K. K. (2020). Adoption of artificial intelligence in higher education: A quantitative analysis using structural equation modelling. *Education and Information Technologies, 25*(5), 3443-3463. https://doi.org/10.1007/s10639-020-10159-7

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). https://doi.org/10.1186/s41239-023-00425-2

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

e-ISBN: 978-629-98082-2-0

Gustilo, L., Ong, E., & Lapinid, M. R. (2024). Algorithmically-driven writing and academic integrity: exploring educators' practices, perceptions, and policies in AI era. *International Journal for Educational Integrity, 20*(1). https://doi.org/10.1007/s40979-024-00153-8

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology, 15*(2), 421. https://doi.org/10.30935/cedtech/13036

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education, 10*(2). https://doi.org/10.1080/2331186x.2023.2236469

Miao, J., Thongprayoon, C., Suppadungsuk, S., Garcia Valencia, O. A., Qureshi, F., & Cheungpasitporn, W. (2023). Ethical dilemmas in using AI for academic writing and an example framework for peer review in Nephrology Academia: A Narrative Review. *Clinics and practice, 14*(1), 89–105. https://doi.org/10.3390/clinpract14010008

Rahman, A., Zulkornain, L. H., & Hamzah, N. H. (2022). Exploring artificial intelligence using automated writing evaluation for writing skills. *Environment- Behaviour Proceedings Journal, 7*(SI9), 547-553.

Robinette, P., Wagner, A. R., & Howard, A. M. (2016). Investigating human-robot trust in emergency scenarios: Methodological lessons learned. In *Springer EBooks*, 143–166. https://doi.org/10.1007/978-1-4899-7668-0_8

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14.

Thomas, L. (2020). Simple Random Sampling | Definition, Steps & Examples. *Scribbr*. https://www.scribbr.com/methodology/simple-random-sampling/

Washington, J. (2023). *The impact of generative artificial intelligence on writer's self-efficacy: A critical literature review.* SSRN. https://doi.org/10.2139/ssrn.4538043

Xu, L., Zhang, J., Ding, Y., Zheng, J., Sun, G., Zhang, W., & Philbin, S. P. (2022). Understanding the role of peer pressure on engineering students' learning behavior: A TPB perspective. *Frontiers in Public Health,* 10. https://doi.org/10.3389/fpubh.2022.1069384

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 1–18. https://doi.org/10.1155/2021/8812542

e-ISBN: 978-629-98082-2-0

P105

From Grammar Checks to Idea Support: Student Insights and Patterns of AI Use in English Assessment at a Malaysian Polytechnic.

***Noor Darliza Binti Mohamad Zamri [1]**

[1]School of Languages, Literacies & Translation, University Sains Malaysia, MALAYSIA.
E-mail: [1]darliza@student.usm.my
**corresponding author: [1]darliza@student.usm.my*

Abstract

This study investigates the use of artificial intelligence (AI) tools in English language assessments within the context of Port Dickson Polytechnic, Malaysia, where the integration of AI into language learning is growing, but its role in assessment remains underexplored. The objective was to examine usage patterns, purposes, perceived fairness, ethical considerations, and policy implications of AI in assessed English tasks. The scope focused on students who were either currently taking or had previously taken an English course at the institution. Data were collected from 120 participants using a survey consisting of seven closed-ended and four open-ended questions. Quantitative data were analysed using descriptive statistics, while qualitative responses underwent thematic coding. Findings revealed diverse usage patterns, with AI most commonly employed for idea generation, grammar/style checking, and translation. Students' views on fairness and usefulness were generally positive, though they also noted concerns about over-reliance, unequal access, and occasional inaccuracies. The study concludes that while AI offers potential to enhance English language assessment, its integration requires clear institutional guidelines, professional development for educators, and policies that address equity and validity. These insights contribute to ongoing discussions on technology in language assessment, offering a context-specific perspective from Malaysian Polytechnics.

Keywords: artificial intelligence; English language testing; language assessment; Malaysian polytechnic; AI usage patterns

1. Introduction

Generative AI tools such as ChatGPT, Grammarly, QuillBot, and DeepL are now embedded in students' day-to-day English work, offering rapid feedback, personalised scaffolding, and productivity gains (Kohnke, Moorhouse, & Zou, 2023). In assessment, however, unresolved questions persist around validity (what construct is being measured when AI assistance is used), fairness/equity (access and bias), and governance (acceptable use and disclosure). Sector guidance increasingly recommends moving beyond prohibition and post hoc detection toward clear task-specific guidance and process-oriented designs, while language-assessment scholars debate where, when, and how assistive technologies should be permissible.

Empirically, students often report using AI for idea support and language repair rather than wholesale text generation—citing efficiency and confidence gains alongside concerns about over-reliance (e.g., Vieriu &

Petrea, 2025). Yet little is known about polytechnic contexts in Malaysia. This study therefore examines how prevalent AI use is, which functions students use most, and how acceptable/fair they judge AI in English assessment at Port Dickson Polytechnic.

Previous research has examined student use of AI in higher education. For example, Vieriu and Petrea (2025) found that Romanian undergraduates used AI for grammar checking, paraphrasing, and idea generation, citing benefits such as personalised learning and improved efficiency, alongside concerns about over-reliance and diminished critical thinking. Despite this growing literature, little is known about how Malaysian polytechnic students use AI in English language assessments. This study addresses this gap by exploring the usage patterns, perceived fairness, and implications of AI use among Port Dickson Polytechnic students.

2. Methods

This study employed a cross-sectional survey design to investigate AI usage in English language assessments among students at Port Dickson Polytechnic. A total of 120 participants were recruited, all of whom were either currently enrolled in or had previously completed an English course at the institution. Participation was voluntary, and respondents were informed that their data would remain anonymous and confidential.

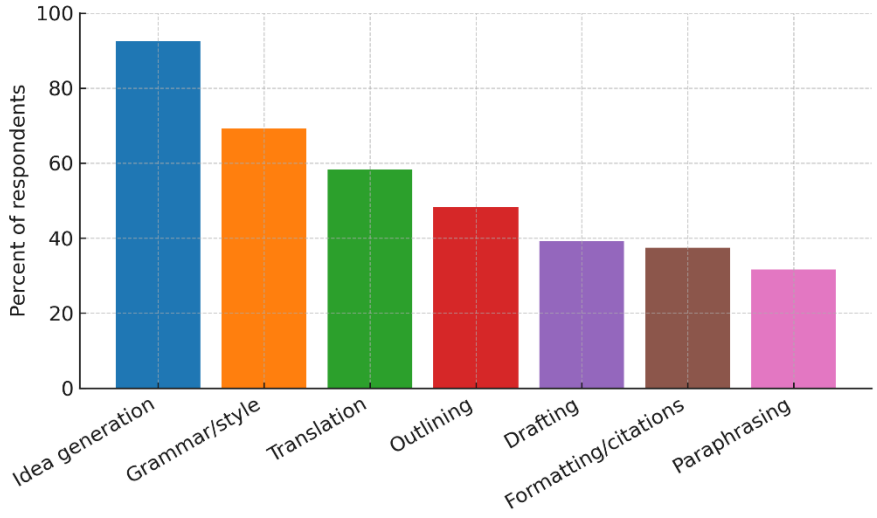
The research instrument consisted of a bilingual questionnaire in English and Bahasa Malaysia to ensure accessibility and comprehension. The survey included seven closed-ended questions, using multiple-choice and Likert-scale formats, and four open-ended questions to elicit qualitative insights. The questionnaire design was informed by the work of Vieriu and Petrea (2025), who investigated AI use in academic contexts through a similar structure focusing on frequency, purpose, benefits, and perceived challenges.

The survey was administered online via Google Forms and distributed through existing student Telegram groups. Respondents could complete the questionnaire at their convenience within a one-week data collection period. Informed consent was obtained at the beginning of the survey, with a clear explanation of the research aims and ethical considerations.

Quantitative data were analysed using descriptive statistics, including frequency counts and percentages, to summarise usage patterns and purposes of AI in English language assessments. Quantitative data were summarised using descriptive statistics (frequencies/percentages). Qualitative responses were analysed via reflexive thematic analysis (Braun & Clarke, 2006). This paper also examined associations between usage and acceptability with simple cross-tabulations (row percentages). For key proportions, the report is N = 120 and, where useful, 95% confidence intervals.

3. Results and Discussion

Figure 1: Purposes of AI use in assessed English tasks

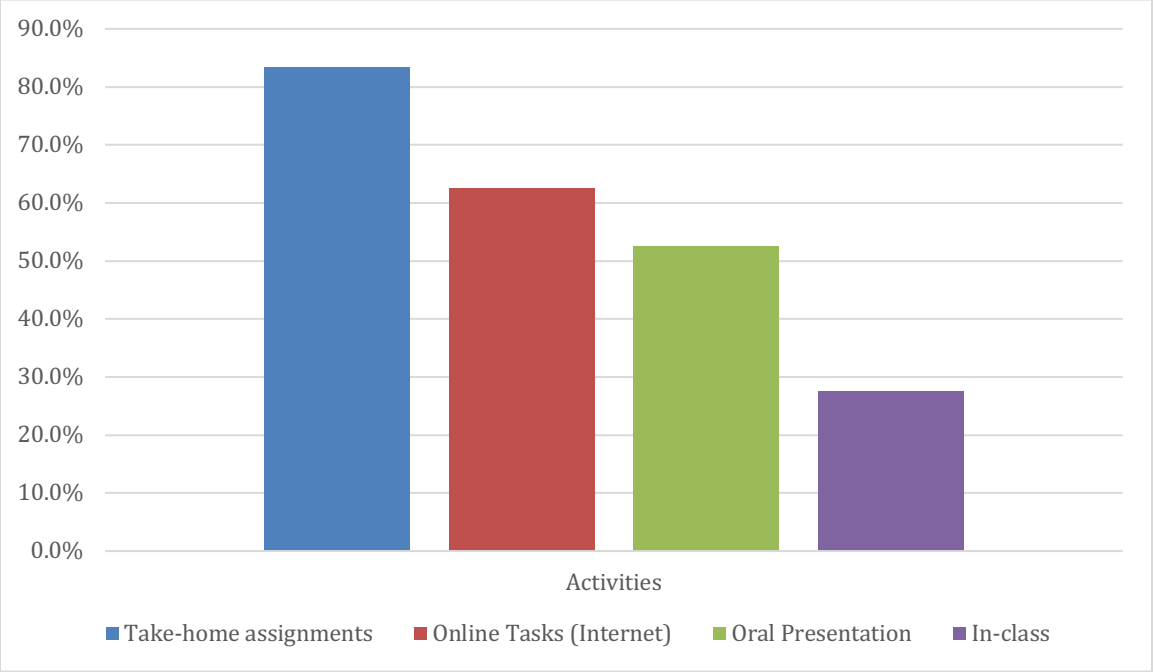


The distribution of functions suggests AI is used primarily as scaffolding rather than as a substitute for authorship. The most frequently reported purpose was idea generation (92.5%), followed by grammar/style checking (69.2%) and translation (58.3%). These high rates point to support at the pre-writing and revision stages—help in overcoming the blank page, refining linguistic form, and mediating comprehension—more than end-to-end text production. Mid-tier uses—outlining (48.3%) and drafting sentences/paragraphs (39.2%)—indicate some movement into content shaping, though still below ideation and repair. Formatting/citations (37.5%) and paraphrasing (31.7%) are present but less central.

This profile aligns with broader HE evidence that students value AI for efficiency, fluency, and confidence, with comparatively fewer reporting wholesale generation of prose. In other words, students appear to be augmenting their own writing rather than outsourcing it. The prominence of translation also reflects a multilingual learning context in which AI is leveraged to clarify meaning before composing or revising.

From an assessment standpoint, the pattern supports process-aware design: permitting AI for explicitly listed functions (e.g., ideation, language repair, reference formatting) in coursework with acknowledgement, while reserving restricted moments (e.g., timed writing/oral tasks) when unaided language control is the construct of interest. Instructors can make this operational by asking for a brief use note or prompt log and by awarding credit for decision-making across drafts (outline → draft → revision). This keeps student agency central, mitigates over-reliance, and preserves validity without ignoring how students actually use AI.

Figure 2: Assessment contexts where students used AI



Reported contexts (multi-select) indicate that AI support is concentrated around assessment rather than within tightly controlled settings. Students most often cited take-home assignments (83.3%), followed by online tasks with internet (62.5%) and oral assessment preparation (52.5%); a smaller share selected in-class tasks without internet (27.5%). Read together with the functions data, this profile is consistent with planning/rehearsal and revision use—generating ideas, organising content, and repairing language—before or alongside coursework submission, rather than sustained, unaided use during closed conditions.

The distribution suggests that guidance and design efforts should prioritise coursework and online task contexts (where most AI use occurs): make permitted functions explicit (e.g., ideation, grammar/style, formatting), require brief acknowledgment of AI assistance, and encourage process evidence (outline/drafts). For in-class, no-internet assessment, the minority selecting this option (27.5%) underscores the need for clear task briefs and allowed resources so expectations are unambiguous, alongside occasional AI-restricted moments when unaided language control is the construct of interest.

Table 1: Open-ended themes and prevalence

Percentages reflect the share of respondents mentioning the theme at least once; blank cells indicate infrequent mentions (<5%) or no exact percentage.

Domain	Theme	Percent of respondents (%)
Benefits	Confidence / support	44.1
Benefits	Idea generation / brainstorming	33.1
Benefits	Language accuracy / grammar	29.7

Benefits	Understanding / translation	14.4
Challenges	Access / connectivity / premium limits	9.3
Challenges	Over-reliance / reduced skills	8.5
Challenges	Incorrect / misleading output	7.6
Challenges	Unnatural language	
Challenges	Integrity / detector anxiety	
Preferred role of AI	Supportive / practice-only role	44.1
Preferred role of AI	Some integration (limited feedback during certain tasks)	10.2
Preferred role of AI	Use with explicit limits & disclosure	
Policy suggestions	Define allowed vs. not-allowed uses per task	13.6
Policy suggestions	Staff / student training	11.9
Policy suggestions	Clear guidelines / policy	11.0
Policy suggestions	Use detection tools (with caution)	≈2.5

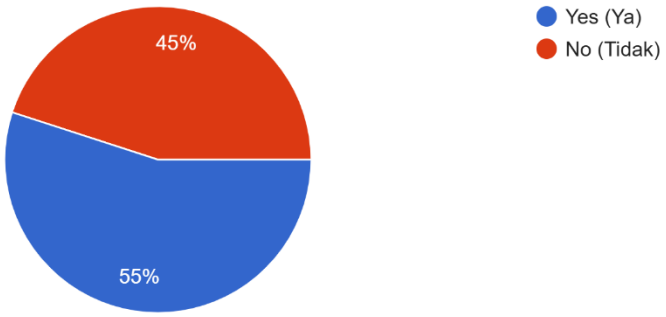
The benefits cluster is dominant and pedagogically coherent: students most often cited confidence/support (44.1%), followed by idea generation/brainstorming (33.1%) and language accuracy/grammar (29.7%). A further 14.4% mentioned understanding/translation. Taken together, these codes position AI as scaffolding for pre-writing and revision—help to get started, to refine form, and to clarify meaning—rather than as a substitute for authorship. This aligns with your closed-item pattern (very high rates for idea generation and grammar/style checking) and with broader HE findings that students value AI for efficiency and fluency gains while retaining ownership of their text.

On the challenges side, the signal is narrower and pragmatic. Respondents pointed to access/connectivity/premium limits (9.3%), over-reliance/reduced skills (8.5%), and incorrect/misleading output (7.6%); smaller mentions referenced unnatural language and integrity/detector anxiety. The emphasis here is less on moral hazard and more on conditions for safe, equitable use (can I access reliable tools? will my skills atrophy? how do I verify output?). This framing mirrors prior studies in which risks are acknowledged but seen as manageable with appropriate guardrails (e.g., verification habits, opportunities to practice unaided writing).

The normative/policy signals are consistent: most respondents favoured a supportive/practice-only role (44.1%), with a smaller group endorsing some integration (e.g., limited feedback during defined tasks; 10.2%). Concrete policy suggestions clustered around defining allowed vs. not-allowed uses per task (13.6%), staff/student training (11.9%), and clear guidelines (11.0%), while detection tools were mentioned only marginally (~2.5%). Read together, these themes point toward process-aware assessment—explicit allowances, brief disclosure of AI assistance, and targeted training—over detector-led policing. In practice, that means specifying permitted functions (e.g., ideation, language repair, formatting) in coursework, requiring a short "AI use" note or prompt log, and reserving AI-restricted moments when unaided language control is the construct of interest.

The discussion of fairness and validity in AI-assisted testing aligns with Bridgeman's (2023) argument that while AI can improve formative assessment, its role in summative evaluation requires careful regulation. In the Malaysian polytechnic context, the absence of clear guidelines leaves both students and educators navigating AI usage informally, which can create inconsistencies in assessment practices. Addressing these challenges will require both policy development and targeted educator training.

Figure 3: Support for allowing AI during testing and subgroup differences



Support for permitting AI during English testing was moderate rather than overwhelming: 55.0% responded *Yes* (95% CI ≈ 46.1–63.9), with the remainder not supportive. Cross-tabulations show that support is not a simple function of use intensity. Among students who reported using AI Always for assessed tasks, only 29.4% favoured allowing it in tests; support was 56.7% for Often, 56.1% for Sometimes, and 75.0% for Rarely. In contrast, normative beliefs about fairness tracked acceptability more closely: 62.7% of those who judged AI use in assessed tasks fair supported allowing it in testing, compared with 11.1% among those who did not judge it fair.

These patterns suggest students differentiate between AI as routine scaffolding in coursework and AI under test conditions, where construct integrity and comparability matter more. The majority "Yes" indicates openness to conditional allowance (e.g., clearly specified functions or tools), while the sizeable "No" signals a preference to protect unaided language control in some tasks. For policy and design, this points to a dual-track approach: (1) declare AI-permitted tasks with explicit limits (e.g., ideation, grammar/style, formatting) and brief disclosure requirements; (2) maintain AI-restricted assessments when unaided writing or speaking is the construct, supported by clear briefs and appropriate invigilation.

The mixed views on allowing AI in testing should also be read alongside concerns about detection practices. While some institutions turn to AI-text detectors, the empirical record shows these tools can misclassify legitimate prose—especially by L2 writers—with error patterns that raise fairness issues (Liang et al., 2023). In our context, where many students write in English as an additional language, over-reliance on detectors risks false accusations and erosion of trust. This further supports a shift toward transparent allowances, disclosure of AI assistance, and process-aware assessment rather than detector-led policing.

4. Conclusion

This study provides empirical evidence on AI use in English assessment at Port Dickson Polytechnic. Usage is already widespread, with students primarily leveraging AI for idea generation, grammar/style, and translation, indicating support across planning and revision stages rather than wholesale text generation. Students generally viewed such use as fair and useful, echoing broader higher-education findings that AI functions as scaffolding to boost efficiency and confidence (Vieriu & Petrea, 2025). At the same time, respondents recognised risks—including over-reliance, unequal access, and occasional inaccuracies—that align with ongoing debates about validity and fairness in assessment (Bridgeman, 2023).

Acceptance of AI in test conditions was moderate, and our cross-tabulations suggest that acceptability tracks normative beliefs about fairness more closely than raw usage frequency. These patterns support a dual-track approach to assessment design: (1) explicitly AI-permitted tasks (e.g., ideation, grammar/style, formatting) with brief disclosure or process evidence (drafts, prompt logs), and (2) AI-restricted tasks when unaided language control is the construct of interest. In line with fairness concerns, institutions should avoid using AI-text detectors as primary evidence of misconduct; if screening is unavoidable, detectors should be used only as triage, with human adjudication, opportunities for students to contest, and triangulation via process artefacts—especially important in L2 contexts given detector bias risks (Liang et al., 2023).

Institutionally, the findings underscore the need for clear, task-specific guidelines, equitable access to approved tools, and professional development so lecturers can both protect construct validity and leverage AI pedagogically. While limited to a single polytechnic, this work motivates multi-site and longitudinal research to examine how sustained AI exposure shapes proficiency, assessment performance, and learner autonomy over time (cf. Vieriu & Petrea, 2025; Bridgeman, 2023).

Acknowledgment

The author is grateful to the participating students and colleagues at Port Dickson Polytechnic for their support in administering the survey. This research received no specific grant from any funding agency.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bridgeman, B. (2023). AI and fairness in educational assessment. *Language Testing, 40*(3), 345–350. https://doi.org/10.1177/02655322231112345

Dizon, G. (2024). A systematic review of Grammarly in L2 English writing research. Cogent Education, 11, 2397882. https://doi.org/10.1080/2331186X.2024.2397882

Educational Testing Service. (2024). Highlights: Responsible use of AI in assessment—ETS principles. https://www.ets.org/research/policy_research_reports/publications/report/2024/kgze.html

Jisc. (2024). Embracing generative AI in assessments: A guided approach [Flowchart]. https://www.jisc.ac.uk/guides/embracing-generative-ai-in-assessments-a-guided-approach

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

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Quality Assurance Agency for Higher Education. (2023). QAA advice and resources on generative AI. https://www.qaa.ac.uk/sector-resources/generative-artificial-intelligence/qaa-advice-and-resources

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

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. Education Sciences, 15(3), 343. https://doi.org/10.3390/educsci15030343

Wang, Y. (2024). Artificial intelligence in second language assessment: Trends and ethical considerations. *Language Assessment Quarterly, 21*(1), 1–20. https://doi.org/10.1080/15434303.2023.2289081