

2026年度大学院博士後期課程入学試験問題

研 究 科 名	科 目 名
文学研究科 英文学専攻	英 語

Doctoral Course English Language Assessment (2026)

For this English language assessment, you are asked to review the five articles provided and draft four or more paragraphs synthesizing key points from the articles, similar to what would be found in a literature review.

Please note the following:

- You do not have to read the complete articles.
- Review the articles, and identify common themes/topics, similar or opposing viewpoints, and/or reported research results.
- Synthesize the key points from all five articles into paragraphs, similar to what would be found in a literature review. Including points from all five articles is acceptable and may improve your examination results.
- Use multiple reference points from multiple articles where possible.
- Follow standard academic essay organization.
- Include a minimum of four paragraphs. Additional paragraphs are acceptable and may improve your examination results.
- Use appropriate APA Style for in-text citations.
- Paraphrase: Do not copy; Do not quote.
- Complete the handwritten assessment within 90 minutes.
- Do not use time to write out a reference list. In-text citations only are all that is required for referencing.

Your assessment will be scored based on the following criteria:

- Level and breadth of synthesis of information from the articles presented
- Relevancy and appropriateness of content presented
- Academic organization
- Language / grammar skills

Articles

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, 33(1), 99–115. <https://doi.org/10.1177/0265532215582283>

González, E. F., Trejo, N. P., & Roux, R. (2017). Assessing EFL university students' writing: A study of score reliability. *Revista Electrónica de Investigación Educativa*, 19(2), 91–103. <https://doi.org/10.24320/redie.2017.19.2.928>

Huang, J., & Whipple, P. B. (2023). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: Implications for educational assessment policy. *Humanities & Social Sciences Communications*, 10(1), 1–10. <https://doi.org/10.1057/s41599-023-02385-4>

Pham, T. N., Lin, M., Trinh, V. Q., & Bui, L. T. P. (2020). Electronic peer feedback, EFL academic writing and reflective thinking: Evidence from a Confucian context. *SAGE Open*, 10(1), 1–20. <https://doi.org/10.1177/2158244020914554>

Yang, Y. (2024). The reliability of using ChatGPT in rating EFL writings. *Shanlax International Journal of Education*, 12(4), 49–59. <https://doi.org/10.34293/education.v12i4.7855>

解答または解答例 :
Sample Answer(s) or Outline:

Literature Review: Systemic Challenges and Innovations in Writing Assessment

Introduction

The assessment of English language writing is a complex psychometric challenge, persistently dogged by issues of fairness and consistency. A consensus across recent scholarship indicates that traditional, single-rater models are increasingly untenable due to the inherent subjectivity of human judgment (González, Trejo, & Roux, 2017; Huang & Whipple, 2023). In response to these reliability deficits, the field is currently undergoing a bifurcation in assessment strategy: one path seeks to repair human scoring through rigorous calibration (Attali, 2016), while another seeks to supplement or replace the human rater entirely through technological and collaborative alternatives (Pham, Lin, Trinh, & Bui, 2020; Yang, 2024).

Point Paragraph 1: The Consensus on Human Unreliability

The most heavily supported theme across the provided texts is the fragility of human scoring reliability. Whether in high-stakes standardized testing or local university classrooms, single-rater assessment frequently fails to meet statistical standards for validity. Huang and Whipple (2023) and González et al. (2017) both provide empirical evidence that professional educators—regardless of their specific location—exhibit significant variability in severity and leniency. While Huang and Whipple (2023) identify this as a failure of generalizability in standardized contexts, González et al. (2017) locate the failure in the inability of shared professional backgrounds to produce consensus. Together, these studies suggest that the "lone teacher" model of assessment is structurally flawed, necessitating either better training or additional evaluative agents to ensure fairness.

Point Paragraph 2: Mitigation Strategies: Calibration vs. Experience

If human raters are inherently unreliable, the literature debates the mechanism for improvement. The conventional assumption that teaching experience equates to assessment reliability is contested. Attali (2016) argues that "calibration"—brief, intensive training with immediate feedback—is the primary driver of accuracy, demonstrating that novices can perform on par with experts when properly trained. This finding complicates the observations of González et al. (2017), who noted discrepancies even among experienced professionals. Synthesizing these views suggests that professional tenure alone is insufficient for reliable scoring; rather, reliability is a dynamic skill that must be actively maintained through specific, data-driven interventions (Attali, 2016), rather than assumed based on a teacher's years of service (González et al., 2017).

Point Paragraph 3: Shifting the Burden: AI and Peer Agents

Given the difficulty of calibrating human teachers, recent research has pivoted toward non-teacher agents to provide feedback and scoring. This shift is evident in the exploration of both artificial intelligence and peer networks as viable evaluators (Pham et al., 2020; Yang, 2024). While the mechanisms differ, the goal is identical: to decentralize the assessment process. Yang (2024) highlights the efficiency of Large Language Models (LLMs) like ChatGPT for scoring language mechanics, though noting they still lack the "human" nuance required for structural critique. Conversely, Pham et al. (2020) utilize Electronic Peer Feedback (e-PF) to enhance reflective thinking.¹ While AI attempts to replicate the teacher's score (Yang, 2024), peer feedback attempts to democratize the process (Pham et al., 2020), suggesting that the future of assessment may lie in a hybrid model that utilizes multiple sources of feedback to overcome the limitations of any single rater.

Conclusionrestate above points

Bulleted Synthesis Points (Multi-Citation Focus)

- **The Universality of Rater Bias:**
- Both standardized testing bodies and classroom teachers struggle to achieve consistent scores when relying on single raters (González et al., 2017; Huang & Whipple, 2023).
- **The "Experience" Paradox:**
- Professional background and experience do not guarantee reliability (González et al., 2017), whereas targeted, short-term training can successfully calibrate even novice raters (Attali, 2016).²
- **Decentralizing Assessment:**
- The field is moving away from the "teacher-as-sole-judge" model, exploring automated systems for efficiency and peer systems for pedagogical reflection (Pham et al., 2020; Yang, 2024).
- **Validity Concerns in New Methods:**
- Just as human raters struggle with consistency, new alternatives face their own validity hurdles, such as AI's inability to judge organization (Yang, 2024) or cultural barriers to peer critique (Pham et al., 2020).

出題意図：

purpose of Questions：

I believe the purpose of the task is clearly outlined so I am copying and pasting it here:

For this English language assessment, you are asked to review the five articles provided and draft four or more paragraphs synthesizing key points from the articles, similar to what would be found in a literature review.

Please note the following:

- You do not have to read the complete articles.
- Review the articles, and identify common themes/topics, similar or opposing viewpoints, and/or reported research results.
- Synthesize the key points from all five articles into paragraphs, similar to what would be found in a literature review. Including points from all five articles is acceptable and may improve your examination results.
- Use multiple reference points from multiple articles where possible.
- Follow standard academic essay organization.
- Include a minimum of four paragraphs. Additional paragraphs are acceptable and may improve your examination results.
- Use appropriate APA Style for in-text citations.
- Paraphrase: Do not copy; Do not quote.
- Complete the handwritten assessment within 90 minutes.
- Do not use time to write out a reference list. In-text citations only are all that is required for referencing.

Your assessment will be scored based on the following criteria:

- Level and breadth of synthesis of information from the articles presented
- Relevancy and appropriateness of content presented
- Academic organization
- Language / grammar skills