Chapter X

Generative AI and the Crisis of Validity in English Language Assessment

Siti Mariam
Universitas Islam Negeri Walisongo Semarang
sitimariam@walisongo.ac.id

1. INTRODUCTION

The rapid deployment of generative artificial intelligence (AI) in English language teaching (ELT) has both delighted and alarmed researchers and educators. Large Language Models (LLMs) can now produce grammatically perfect documents, provide instant feedback, and produce model responses for speaking and writing activities. Teachers' lesson plans and students' language engagement are being altered by the increasing integration of ChatGPT and other comparable tools into educational environments. While these technologies offer increased efficiency and accessibility, they also pose serious questions about how algorithmic systems represent and reproduce linguistic norms. Critical study shows that algorithms always replicate the biases and hierarchies present in the training data, despite the fact that AI-generated feedback is often perceived as neutral and objective.

Sociolinguistics and practical linguistics have long acknowledged the global spread of English as a politically and culturally complicated phenomenon. The concept of World Englishes highlights the existence of several English dialects in Inner-Circle, Outer-Circle, and Expanding-Circle scenarios. Despite increased awareness of linguistic diversity, standardized Inner-Circle norms, particularly British and American English, continue to dominate language evaluation, academic writing, and ELT instruction (Rose, H., et al.,

2021). In addition to being linguistic conventions, these standards function as symbols of authority that restrict access to career and educational prospects.

Because of this, language education has sometimes been charged with maintaining global linguistic hierarchies by giving certain English dialects more authority and legitimacy. The emergence of generative AI adds a new dimension to these long-running discussions. Since large language models are trained on massive digital corpora that are dominated by texts written in economically and technologically advanced places, they may implicitly favour Inner-Circle linguistic patterns. When AI systems generate model texts or provide corrective feedback, they may normalize some types while marginalizing others, thereby changing learners' perceptions of what constitutes "correct" or "appropriate" English.

2.DISCUSSION

This raises an important question for ELT research: Do generative AI systems reinforce Inner-Circle English norms, hence sustaining language inequality in digital form? To answer this question, generative AI must be seen not only as a teaching tool but also as a sociotechnical system interwoven into global systems of language power and knowledge production.

2.1. Theoretical Framework: Validity in Language Assessment (Messick, Kane, Bachman & Palmer)

Validity is often recognized as the most important element in educational and language assessment. Several notable researchers, including Samuel Messick, Michael Kane, Lyle Bachman, and Adrian Palmer, have influenced how validity is conceptualized in modern testing theory. Together, their frameworks provide as a theoretical platform for investigating how Generative AI challenges the interpretation and application of language assessment scores. Messick's unified theory of validity defines validity as the extent to which data and theoretical rationales support test score interpretations and applications. Rather than addressing validity as distinct categories (e.g., content validity or criterion validity), Messick

advocates for a unitary concept that incorporates numerous sources of evidence, such as content, substantive, structural, generalizability, external, and consequential features. Within this approach, a primary consideration is whether a test accurately measures the desired construct. The rise of Generative AI compounds this issue because AI-assisted language output may introduce construct-irrelevant variance, resulting in test scores that reflect technological aid rather than the learner's underlying language competency. Kane builds on Messick's work by proposing an argument-based approach to validation, emphasizing the significance of properly expressing the interpretive chain that connects observable performance to score-based conclusions. Kane defines validation as the construction and evaluation of an interpretative argument made up of various inferences, such as scoring, generalization, extrapolation, and decision-making. In the setting of AI-mediated language generation, these assumptions become unstable. These statements are supported by (Bachman, L.F., & Palmer, 2010; Kane, 2013; Messick, 1996).

For example, the extrapolation inference—which links test performance to real-world language ability—may be diminished if test results are heavily influenced by AI tools rather than the learner's own linguistic resources. Bachman and Palmer (2010) broaden validity discussions with their model of communicative language capacity and assessment usefulness. Their concept focuses on the relationship of linguistic ability, test task features, and test takers' strategic competency. Many factors influence assessment usefulness, including reliability, construct validity, authenticity, inter-activeness, effect, and practicality. Generative AI calls into question several of these traits, particularly authenticity and inter-activeness, because test performance may increasingly require human-AI interaction rather than merely individual language output.

Together, these theoretical approaches demonstrate how the advent of Generative AI poses new challenges to existing language assessment paradigms. By challenging whether test results actually represent individual communicative competence, AI emphasizes the need to re-examine the concept of validity in modern language testing environments.

2.2. Validity Theory Revisited: From Measurement to Interpretation

The degree to which test score interpretations for intended purposes are supported by theory and evidence is the traditional definition of validity. This idea holds that a test's validity is not decided by the test itself, but rather by the conclusions that can be drawn from it. The necessity of defining and evaluating the line of reasoning that links claims about ability to observable performance is highlighted by (Kane, 2013) argument-based method. However, generative AI complicates this inferential chain. When AI systems are used to generate text, algorithmic assistance mediates the relationship between observed performance and underlying competency. This mediation challenges the validity of ideas such as "writing proficiency". If competency has traditionally been characterized as the ability to independently construct coherent discourse, then AI co-production challenges concept representation. Rather than only technical contamination, the issue is conceptual drift. As (Bachman, L.F., & Palmer, 2010; Kane, 2013; Mariam, S., et al., 2025b) notes, assessment theories must evolve together with communicative practices. If real-world literacy increasingly incorporates digital tools and AI-supported composition, excluding such mediation from assessment could render exams environmentally meaningless.

Additionally, (Messick, 1996) emphasis on consequential validity stands apart. Decisions about assessments have societal consequences, influencing career and educational choices. Punitive detection-based strategies may disproportionately impact students with inadequate digital literacy in AI-mediated environments, neglecting underlying conceptual issues. As a result, the subject is how to reframe validity arguments to incorporate hybrid forms of performance rather than how to exclude AI from assessment.

A dynamic competency model may offer a viable path. Competence could be understood as the ability to strategically manage language and technological resources, rather than being synonymous with autonomous innovation. This shift is consistent with interactional and socio-cognitive models of language ability, which see proficiency as relational, contextual, and context-dependent (Ibrahim, 2023; Mariam, S., et al., 2025; Mizumoto, A., & Eguchi,

2023). According to this viewpoint, validity must take into account the performance ecology, which includes technological mediation.

2.3. Epistemic Authority and Distributed Cognition

By changing power over knowledge creation, generative AI challenges the epistemic foundations of language evaluation. Epistemic authority refers to the legitimacy that particular individuals or institutions have in establishing what constitutes knowledge. This influence has traditionally been held by educational institutions and test developers, who establish the standards for linguistic proficiency and correctness. However, when compared to unskilled writers, algorithmic systems trained on large corpuses generate works that more consistently adhere to institutional standards.

The concept of distributed cognition serves as a useful theoretical foundation. Cognitive processes are distributed across individuals, things, and surroundings rather of being restricted to the individual mind. This approach holds that AI systems are cognitive extensions rather than outer invasions. In a similar spirit, (Clark, A., & Chalmers, 1998) "extended mind" theory claims that tools can evolve into critical components of cognitive systems. If this is the case, perceiving AI-aided writing as fundamentally wrong may reflect an old individualist epistemology.

However, algorithmic power creates issues of bias and transparency. AI models are trained using large datasets that reflect current linguistic norms, and they frequently support Inner-Circle versions of English. This risk marginalizing distinct varieties of English and establishing linguistic hierarchy. Assessment systems may unintentionally prolong inequality by unequivocally elevating AI-generated norms. Tests are weapons of power and policy, and their development and interpretation have ideological implications, as (Johnston, H., et al., 2024; Mariam, S., et al., 2025; Uto, M., & Aramaki, 2024) points out. As a result, while redefining validity, epistemic justice must be considered. Assessment frameworks must be used to investigate the legitimacy of knowledge as well as how algorithmic systems influence acceptable standards. A balanced paradigm that includes both technical aid and human expertise, rather than replacing human judgment with automated scoring, could bring greater responsibility.

2.4. Authorship, Agency, and Posthuman Perspectives

Because written texts are supposed to indicate individual intellectual ownership, authorship has long been important in academic evaluation. This idea is challenged by post-humanist and socio-materialist perspectives, which regard agency as coming from relational networks rather than isolated individuals (Hannah, L., et al., 2023; Lin, Z., & Chen, 2024; Mariam, S., et al., 2025a). Because generative AI actively influences textual production, writing in educational settings has always been tempered by tools, feedback, and collaborative resources. As a result, it may be important to reframe language proficiency as strategic competence, or the ability to critically manage linguistic and technological resources. This viewpoint promotes process-oriented evaluation, which stresses reflective and iterative writing techniques and is consistent with sociocultural mediation and scaffolding theories (Aryadoust, V., et al., 2024; Mariam, S., et al., 2018; Mesgar, M., & Strube, 2024).

2.5. Conceptual Structure of AI-Mediated Language Assessment

Language performance is rethought as the result of an interaction between human linguistic ability, AI mediation, and assessment task conditions in the suggested conceptual framework of AI-mediated language assessment. The existence of generative AI adds an extra mediating layer that modifies how language is created and assessed, in contrast to conventional language assessment models that presume that test results directly reflect a learner's communicative competence. Human linguistic competence, which is based on the idea of Communicative Language Ability (CLA) put forth by Bachman and Palmer, forms the basis of the model. Grammar knowledge, textual structure, pragmatic appropriateness, and strategic competence are all included in this competency. These qualities are thought to directly produce visible language performance in traditional testing scenarios.

However, in AI-mediated contexts, the relationship between competence and performance becomes more complicated, as

learners may rely on AI tools to help generate or refine language output. The model's second component is AI mediation, which depicts the impact of generative AI systems on the language development process. AI tools can help with grammatical corrections, lexical augmentation, content creation, and discourse structuring. As a result, the created language may represent both the learner's own linguistic resources and the linguistic patterns encoded in AI models. This results in a type of algorithmic co-authorship in which human purpose and machine-generated language both impact the final work.

The third component is assessment task conditions, which govern the degree to which AI mediation happens during the assessment. Task designs can range from AI-restricted environments, where technology is not permitted, to AI-aware or AI-integrated activities, which openly incorporate AI tools into the evaluation process. These task settings have a substantial impact on how learners interact with language and artificial intelligence systems. Within this framework, observable language performance results from the interaction of human competency, AI mediation, and task design. As a result, while interpreting evaluation scores, it is necessary to examine how much AI aid effects performance. This interaction raises serious concerns regarding the validity of language assessments, as results may represent not only verbal competency but also learners' capacity to effectively communicate with AI technologies.

3.FUTURE DIRECTIONS

Generative AI requires a fundamental rethinking of validity in English language evaluation. Rather than being viewed as a danger to academic integrity, AI should be interpreted as showing difficulties within traditional assessment methods based on individual cognition and independent authorship. Validity must thus be examined in the context of dispersed intelligence and hybrid human-AI creation. Assessment frameworks should include ecological authenticity, epistemic openness, and ethical accountability while maintaining justice and equity. As epistemic authority shifts to algorithmic systems, institutions must critically evaluate the linguistic norms encoded in AI technology. Future study should look into AI-integrated

assessment models that balance technological innovation, language diversity, and pedagogical integrity.

Author

	Dr. Siti Mariam, M.Pd is a lecturer, researcher and trainer in education, in the Faculty of Education and Teacher Training, Universitas Islam Negeri Walisongo Semarang, Indonesia. Her research interests include Language Testing and Assessment, TESOL Methodology, Curriculum and Material Development. Her email address is sitimariam@walisongo.ac.id. Orcid ID: https://orcid.org/0000-0002-3639-1301. Scopus ID: 59790939700. Sinta ID: 5972930, and Google Scholar.

References

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the Affordances of OpenAI's Large Language Model in Developing Listening Assessments. *Computers & Education: Artificial Intelligence*, *6*. https://doi.org/10.1016/j.caeai.2024.100204

Bachman, L.F., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1111/1467-8284.00096

Hannah, L., Jang, E.E., Shah, M., & Gupta, V. (2023). Validity Arguments for Automated Essay Scoring of Young Students' Writing Traits. *Language Assessment Quartely*, *20*(4–5), 399–420. https://doi.org/10.1080/15434303.2023.2288253

Ibrahim, K. (2023). Using AI-based Detectors to Control AI-Assisted Plagiarism in ESL Writing: "The Terminator Versus the Machines." *Language Testing in Asia*, *13*. https://doi.org/10.1186/s40468-023-00260-2

Johnston, H., Wells, R.F., Shanks, E.M., Boey, T., & Parsons, B. (2024). Student Perspectives on the Use of Generative Artificial Intelligence Technologies in Higher Education. *International Journal for Educational Integrity*. https://doi.org/10.1007/s40979-024-00149-4

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Lin, Z., & Chen, H. (2024). Investigating the Capability of ChatGPT for Generating Multiple Choice Reading Comprehension Items. *System*, *12*(3). https://doi.org/10.1016/j.system.2024.103344

Mariam, S., Fadlilah, S., Kepirianto, C. (2025a). Online Assessment of English Competence and Its Washback: Teachers' and Students' Voices. *JSL: International Journal of Social Learning*, *5*(3), 574–589. https://doi.org/10.47134/ijsl.v5i3.437

Mariam, S., Fadlilah, S., Kepirianto, C. (2025b). What Are the Washback Effects of Online Students' English Competence Assessment? *AWEJ: Arab World English Journal*, *16*(2), 445–458. https://dx.doi.org/10.24093/awej/vol16n02.25

Mariam, S., Kepirianto, C., Fadlilah, S. (2025). Indonesian EFL Teachers' Beliefs and Practices in the Implementation of Authentic Assessment for Speaking Skills. *Forum for Linguistic Studies*, *07*(12), 1290–1301. https://doi.org/10.30564/fls.v7i12.11738

Mariam, S., Saleh, M., Warsono., Mujiyanto, J. (2018). Using the Rasch Model for the Affective Assessment of EFL Learners. *AWEJ: Arab World English Journal*, *9*(2), 441–456. https://dx.doi.org/10.24093/awej/vol9no.2.29

Mesgar, M., & Strube, M. (2024). A Survey on Deep Learning-based Automated Essay Scoring and Feedback Generation. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-11017-5

Messick, S. (1996). *Validity of Performance Assessments* (G. W. Phillips (ed.); I issues i). National Center for Education Statistics.

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). https://doi.org/10.1016/j.rmal.2023.100050

Rose, H., Mc.Kinley, J., Galloway, N. (2021). Global Englishes and Language Teaching: A Review of Pedagogical Research. *Language Teaching*, *54*(2), 157–189. https://doi.org/10.1017/S0261444820000518

Uto, M., & Aramaki, K. (2024). Linking Essay Writing Tests Using Many Facet Models and Neural Automated Essay Scoring. *Behavior Research Methods*, *56*, 8450–8479. https://doi.org/10.3758/s13428-024-02485-2