Annual Review of Applied Linguistics (2025), **45**, 200–221
doi:10.1017/S0267190525000078

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Revisiting communicative competence in the age of AI: Implications for large-scale testing

Xiaoming Xi

Hong Kong Examinations and Assessment Authority, Hong Kong SAR
Email: xxi@hkeaa.edu.hk

Abstract

Changes in the characterization of communicative competence, especially in the context of large-scale testing, are typically driven by an evolving understanding of real-world communication and advancements in test construct theories. Recent advances in AI technology have fundamentally altered the way language users communicate and interact, prompting a reassessment of how communicative competence is defined and how language tests are constructed.

In response to these significant changes, an AI-mediated interactionalist approach is proposed to expand communicative competence. This approach advocates for extending the traditional concept of communicative competence to encompass AI digital literacy skills and broadened cognitive and linguistic capabilities. These skills enable effective AI tool usage, as well as the interpretation and application of AI-generated outputs and feedback, to improve communication. Embedding these competencies into language assessments ensures alignment with contemporary communication dynamics, enhancing the relevance of language assessments, and preparing learners for navigating AI-augmented communication environments.

While high-stakes testing faces considerable challenges in adopting this expanded construct, low-stakes formative assessments, where scores do not influence critical decisions about individuals and where opportunities exist to rectify errors in score-based actions, if any, provide a fertile ground for exploring the integration of AI tools into assessments. In these contexts, educators can explore giving learners access to various AI tools, such as editing and generative tools, to enhance assessment practices. These explorations can start to address some of the conceptual challenges involved in applying this expanded construct definition in high-stakes environments and contribute to resolving practical issues.

Keywords: definition of communicative competence; generative AI; construct; large-scale testing

Introduction

From the time technology was perceived as irrelevant to the measurement of language ability (Taylor et al., 1998) to the current advocacy for integrating generative AI tools into the language construct, or the skills measured by language tests (Voss et al., 2023; Xi, 2023), the landscape of language testing has undergone vast transformations in the

last two decades, driven by recent accelerated technological advancements. Despite intense debates regarding the nature of language constructs and speculations about the future of language testing, large-scale, high-stakes testing practices have remained largely unchanged other than the introduction of integrated tasks starting from the early 2000s (see later discussion of this).

Language tests are expected to closely mimic and elicit real-world communication. The recent widespread use of assistive and generative AI tools by language users in everyday life has prompted new inquiries into the nature of communicative competence, urging language testers to revisit the concept of communicative competence and explore how testing practices can be transformed.

In this article, I will examine the driving forces behind the impetus to redefine communicative competence in relation to the use of AI technology by real-world language users. Guided by a conceptual definition, I will discuss how language assessment tasks might be redesigned to evaluate the broadened constructs. While acknowledging the significant potential for introducing innovations in low-stakes formative assessments to reflect changes in real-world communication, I will highlight the conceptual and practical challenges of operationalizing the expanded constructs in large-scale high-stakes testing. Finally, I will propose a pathway for introducing incremental innovations in high-stakes contexts.

Factors affecting the definition of language ability for assessment/testing

Language testers address five primary questions: why measure, what to measure, how to measure it, how to report the measurement results, and how to use the results. The first question addresses the paramount issue of test purpose and use, that is, why do we need to use a language test in a particular context? How are the scores going to be used? The second question, what to measure, pertains to the conceptualization of language ability and forms the foundation of language testing. The third question addresses the operationalization of language ability measurement through the design and development of test items and tasks as well as scoring rules and rubrics. The fourth question concerns the aggregation and reporting of test results and communication of the results to users, whereas the final one involves the use of test results and the related consequences, whether intended or unintended.

The construct being measured has always been central to language testing. The definition of language ability or communicative competence influences how language is learned, taught, and assessed. Over the past few decades, large-scale testing, most notably English language proficiency testing, has played a significant role in shaping local classroom practices including learning and instructional methods and emphases. Its influence is evidenced by extensive washback research, which demonstrates how high-stakes tests such as the Test of English as a Foreign Language (TOEFL), the International English Language Test System (IELTS), the College English Test Band 4 and Band 6 in China, and the Center Test in Japan have driven changes in local learning, teaching, and assessment practices (Alderson & Hamp-Lyons, 1996; Green, 2003; Jin, 2000; Read & Hayes, 2003; Saville & Hawkey, 2004; Wall & Horak, 2006; Watanabe, 1996).

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

In principle, what is assessed in language tests should be driven by the nature and characteristics of real-world language communication. However, the connection between language constructs and real-world communication may not be as direct as initially assumed (Leung, 2022; Xi et al., 2021). It must first be articulated through test theories, which aim to define how language ability should be conceptualized for the purpose of designing language tests. These theories, examined in the next section, provide guidance to practitioners in developing language tests and form the basis of standards and frameworks upon which language tests can be designed.

Moreover, the design of any test, particularly large-scale tests, is subject to practical constraints. These constraints, which include test delivery, administration, psychometrics, development, and scoring, significantly impact test practices. The operationalization of language ability in test design is always influenced by these constraints, necessitating a reduction in operational complexity to a manageable level and containment of costs to an acceptable level, while prioritizing the value provided to test users. Consequently, mimicking real-world communication in test design, especially in large-scale testing, is often challenging due to cost, practicality, lack of technological resources, and other concerns. What eventually gets operationalized is influenced by the specific test construct theory employed and is subsequently simplified, reduced, or modified to accommodate various practical constraints.

Conceptual approaches to defining language ability

In this section, four major conceptual approaches to defining language ability are reviewed to provide theoretical grounding for construct definitions: task-based, trait-based, interactionalist, and psycholinguistic approaches. The term, language ability, is used here because only two of the approaches are explicitly rooted in the concept of communicative competence. While abilities to communicate in the real world are the ultimate focus of inferences about language ability drawn from performance on language tests, these approaches concentrate on *how* to characterize underlying language abilities that support such real-world communication.

The task-based approach

Initiated as an approach to encourage teaching of authentic language use (task-based language teaching), the task-based approach conceptualizes the construct of language as the performance on specific tasks (Ellis, 2003; Long, 2015; Skehan, 1998). Norris (2016, p. 232) defines task-based language assessment as "the elicitation and evaluation of language use (across all modalities) for expressing and interpreting meaning, within a well-defined communicative context (and audience), for a clear purpose, toward a valued goal or outcome." This definition highlights the importance of assessing language skills in a specific context and for a particular audience, ensuring that the communication serves a clear purpose and contributes toward a meaningful outcome.

The task-based approach emphasizes the incorporation of genuine communicative tasks into classroom instruction. In test design, the focus is on employing tasks that

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

replicate real-world language activities, with the objective of aligning the characteristics of test tasks with those encountered in authentic scenarios. However, due to various practical constraints such as test length, cost, and the complexity of operations, it is often impractical to encompass all essential tasks within the domain and to fully replicate real-world language activities in the design of large-scale tests. As a result, the ability to make inferences about learners' capability to handle real-world communication demands based on task performance would be compromised (Bachman, 2007). Relying solely on the task-based approach thus poses challenges in establishing a robust conceptual link between test task performance and real-world language proficiency.

The trait-based approach

The trait-based approach centers on the trait (i.e., the language knowledge components, such as grammatical or vocabulary knowledge) as the primary focus, which accounts for performance in the language use domain (Carroll, 1961; Lado, 1961). In test design, the priority is to measure the trait that accounts for consistency in performance across a variety of tasks, with the trait serving as the link between test performance and domain performance.

The trait-based approach endeavors to define the underlying language knowledge, skills, and abilities that account for a learner's performance on both test and real-world tasks, thus elevating language competence to an abstract level. However, the assumption that language ability remains consistent across various contexts has been challenged both theoretically and practically. Advocates of the interactionalist approach, including Chapelle (1998), Chalhoub-Deville (2003), and Xi et al. (2021), dispute the notion of uniform language ability, asserting that language competencies may vary depending on the context of use. The existence of tests for specific purposes, such as TOEFL, IELTS, the Occupational English Test, and the Cambridge English Qualifications Business (BEC), designed for academic and workforce contexts, underscores the understanding that language ability is context-dependent.

Individual variations in language competencies across different contexts are primarily influenced by prior exposure, learning priorities, or a combination of both. Exposure to varied linguistic environments and differing learning priorities uniquely shape an individual's communicative competences. This highlights the need for a more dynamic and context-sensitive approach to understanding language competence. A static, trait-based approach, focusing on inherent language traits, may not adequately account for these variations in diverse language use contexts and not fully capture the richness and fluidity of language use across different contexts.

The interactionalist approach

The interactionalist approach acknowledges the significant role that context plays in the construct of language ability. It posits that certain components of language traits, such as vocabulary and discourse competence, are more susceptible to contextual variations. In contrast, elements like pronunciation may exhibit greater stability (Xi, 2015; Xi et al., 2021). Rather than explaining performance irrespective of context, the construct accounts for consistency in performance across a range of similar contexts.

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

Consequently, inferences about test takers' abilities should be confined to a set of specific contexts.

Recent theoretical frameworks have aligned with the interactional perspective, though they differ somewhat regarding the role of context in defining constructs (Chalhoub-Deville, 2003; Chapelle, 1998; Xi et al., 2021). These approaches vary in how they characterize the interaction between language use contexts and the linguistic resources that individuals bring to communicative situations, as well as how these characterizations influence interpretations of language ability.

Chapelle's (1998) interactionalist approach, labeled as minimalist interactionalist by Bachman (2007), primarily caters to the interests of language testers. Rooted in the perspective that language tests aim to provide stable inferences about individuals' language ability, Chapelle explains that the interpretation of language ability hinges on the consistency of performance across contexts. She highlights the significance of strategic competence or metacognitive strategies (Bachman, 1990; Bachman & Palmer, 1996) in managing the application of linguistic knowledge and skills within specific language use contexts.

Chalhoub-Deville (2003), drawing on the interactionalist literature in second language learning (Young, 2008), conceptualizes the construct as "ability-in-individual-in-context" (p. 372). She argues that the abilities activated in a specific text both influence and are influenced by that context. This perspective extends beyond Chapelle's (1998) approach by recognizing the reciprocal and dynamic nature of context and ability, which can change from moment to moment. This conceptualization offers a richer and more intricate representation of the complex interactions between context and ability, particularly relevant to second language learning researchers. However, in operationalizing this perspective for test design, particularly for large-scale tests, it may be very challenging to adequately capture the nuances of this approach.

Xi (2015) and Xi et al. (2021) further develop this interactionalist approach by introducing a framework designed to characterize contexts of language use for language test design, which are defined in layers such as sub-domains, medium of communication, the communicative setting, etc. She argues that in designing language tests, it is not realistic to represent all these layers due to practicality concerns. Instead, she recommends prioritizing the layers that are believed to most significantly influence the demonstration of language abilities in a specific domain. Additionally, she identifies components of language competence that are most affected by contextual variations, providing practical guidance for implementing this approach in test design.

This interactionalist approach considers the interaction between specific traits, contextual factors, and the dynamic nature of language use, allowing for a more nuanced assessment of language skills. This perspective aligns with current trends in language assessment, which emphasize the importance of assessing real-world communication skills in a target domain. By integrating context-sensitive elements into language constructs, we can better evaluate an individual's ability to navigate specific communicative situations, thus offering a more relevant measure of language competence. Despite their differences, the three approaches reviewed above share the fundamental premise that test constructs must account for performance consistency across a range of similar contexts and that the interpretation of language ability is dependent on the contexts of use.

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

The psycholinguistic approach

The psycholinguistic approach to defining language ability emphasizes processing competence (Van Moere, 2012), which has also been narrowly defined as automaticity in language processing (Figueiredo, 2021; Hulstijn, 2011). Van Moere (2012) expands this definition to include fluency, accuracy, and complexity – three key indicators of language quality that have been extensively researched in second language acquisition and learning (Skehan, 1998).

Researchers in this field argue that in addition to higher-order skills such as interactional competence, pragmatic competence, organization, and coherence, psycholinguistic constructs, such as fluency or automaticity of speech production, lexical access, phonology, and syntactic production, are strong predictors of language ability and should, therefore, be assessed in language tests. This approach highlights the processing components of language ability and emphasizes the importance of psycholinguistic constructs, which are particularly useful for guiding assessment practitioners in designing diagnostic assessments to identify the underlying skills that contribute to deficiencies in higher-order competencies.

However, applying this approach to construct definition in the context of large-scale, high-stakes testing, where efficient test item types are predominantly used without adequately representing communicative tasks, could significantly underrepresent the construct of interest and have potential negative impact on teaching and learning.

Expanding language constructs in relation to AI technology use

When significant breakthroughs occur in the characterization of real-world communicative competence, due to changes in real-world communication and/or advancements in construct theories, it becomes imperative to reconsider how we define communicative competence for assessment.

Over the past 2 years, generative AI technology has significantly changed the way language users leverage AI tools to enhance communication skills, prompting a re-evaluation of the concept of communicative competence. Generative AI refers to systems with the capability to generate extended, human-like text, images, and/or other forms of media using generative models trained on massive amounts of data (Sengar et al., 2024). Prior to the emergence of ChatGPT in late 2022, in the field of language learning, the range of tools accessible to learners was confined to those supporting feedback on primarily linguistic conventions, particularly grammar and vocabulary, such as Grammarly, Criterion, and Write & Improve. With generative AI tools like ChatGPT, however, users can now generate extended writing or speech tailored to their specific communication needs. Additionally, it offers instant and comprehensive feedback on vocabulary, grammar, content, organization, coherence, audience awareness, etc. This transformative shift highlights the expanding role of digital AI capabilities in facilitating real-world communication.

Historical perspectives on the role of technology in language constructs

This section examines the impact of technology on the way language ability is defined and measured. Historically, the content and format of large-scale language tests have

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

closely mirrored advancements in language learning and teaching methodologies. From the 1950s to the 1980s, various teaching methods, such as the grammar-translation method, the direct method, and the audio-lingual method, were prominent. Until 1979, TOEFL assessed reading, listening, structure, and written expressions. With the rise of communicative language teaching (CLT) in the late 1970s and 1980s, which has continued to gain traction, the Test of Spoken English was introduced in 1979 as part of the TOEFL portfolio, followed by the Test of Written English in 1986. The English Language Testing Service, later rebranded as the IELTS, debuted in 1989 as a language test for admissions that includes all four skills of reading, listening, speaking, and writing. In 2005, the TOEFL internet-based test was completely redesigned to include authentic language use tasks that assess all four skills and the ability to use multiple language skills for communication. This redesign drew on the task-based teaching approach, a method within the broader framework of CLT that has gained popularity since the 1990s.

The introduction of integrated tasks that engage multiple modalities marked a significant innovation in large-scale testing in the early 2000s. This approach has been incorporated into large-scale tests such as the TOEFL Internet-Based Test (TOEFL iBT), the Pearson Test of English (PTE), and the Hong Kong Diploma of Secondary Education (HKDSE) English Language test. The HKDSE English Language test, for instance, includes three integrated listening and writing tasks alongside two stand-alone writing tasks. Additionally, the HKDSE employs a group discussion task for the speaking section, which simulates common real-world communicative activities. Furthermore, the concept of English as a lingua franca has increasingly influenced the design of large-scale testing. This began with the IELTS listening test incorporating multiple first language English varieties, followed by the TOEFL iBT adding the British and Australian accents into the listening section, and more recently, PTE Academic test including both first and second language varieties of English in its listening section. These developments, along with the use of integrated tasks, have been driven by language testers' focus on real-world academic language use, where multiple modalities – reading, listening, speaking, and writing – are used simultaneously, and second-language users frequently interact with speakers of diverse English varieties.

However, the influence of technology, aside from its use in facilitating test delivery (e.g., TOEFL iBT, PTE Academic, and the Duolingo English Test), has not been evident in the design of test tasks in large-scale testing. On the contrary, 25 years ago, technology was clearly treated as a source of construct-irrelevant variance. In the 1990s and 2000s, numerous studies, including those for TOEFL (Taylor et al., 1999) and IELTS (Chan et al., 2017; Green & Maycock, 2004), investigated the impact of computer familiarity on test performance and sought to establish validity evidence that the measurement of language ability was not "contaminated" by test takers' computer proficiency. This strong proposition persisted as the mainstream view and remained largely unchallenged until the significant emergence of generative AI applications in late 2022. Researchers have since begun to query the implications for the definition of communicative competence given remarkable changes in the ways learners communicate and interact via written or spoken means. Suddenly, the boundaries of this long-standing

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

comfort zone for language testers, which had confined their work for decades, began to shift or expand to adapt to this drastic change in real-world communication.

Today, editing and generative AI tools are being considered for integration into the construct of communicative competence by providing learners with, initially, guided and monitored access to these tools during assessments (Voss et al., 2023; Xi, 2023). The conceptions of language tests have shifted beyond imagination within just around 25 years due to accelerated AI technology breakthroughs.

Despite the significant advancements in integrating AI tools into local language assessments, large-scale tests have yet to reflect the growing influence of AI technology on real-world communication. For instance, spelling and grammar checkers are routinely used in everyday writing, and now generative AI tools are increasingly utilized to enhance real-world writing outputs. However, even today, none of the large-scale tests allow the use of assistive editing tools by test takers when taking a writing test, not to mention generative AI tools.

Applications of technology in the assessment process

To explore the integration of AI technology into language constructs, it is crucial to differentiate between technology as a medium or aid for real-life communication and as a tool for test delivery or creating interactive test tasks. Historically, these roles may have often been conflated when discussing the role of technology in language constructs. Developments in the former, such as the use of Zoom for online communication, spelling and grammar checkers, online dictionaries, generative AI tools, and real-time translation tools, compel a reevaluation of the construct of language communication, as their use fundamentally challenges our conventional definition of communicative competence.

Conversely, when technology is used to enhance test delivery, such as in delivering computer-delivered monologic speaking tests, the primary goal is to standardize and scale the assessment of speaking skills efficiently, rather than replicating real-world communication such as in Zoom-based communication. In communication delivered by Zoom or other video conferencing technology, aspects of interactional competence, such as turn-taking, may be affected by the absence of visual cues signaling an intention to maintain, yield, or take the floor. Thus, Zoom-based speaking tests attempt to mimic one type of real-world communication scenarios, if the construct is defined as the ability to handle online interactions.

Like the use of technology to deliver monologic speaking tests, when chatbot technology is employed to create interactive speaking or writing tasks, the aim is to simulate real-world communication by replacing the human interlocutor with a chatbot, thus enabling the scaling of interactive assessments that would otherwise be costly and time-consuming to operationalize.

Despite these technological enhancements, the fundamental nature of the real-world language ability, which language tests purport to measure, remains unchanged, thus not impacting the construct definition. This distinction is crucial as we consider the precise role technology plays in defining language test constructs.

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

The role of technology in construct definition

Expanding on Xi et al. (2024), four approaches to addressing the role of technology in defining language test constructs are described: outright rejection, forced acceptance, cautious acceptance, and progressive embracing. The first approach, outright rejection, views technology as a potential source of construct irrelevance. For example, within this approach, using a paper-and-pencil format to test writing skills is based on the belief that writing conventions such as spelling and handwriting are core components of writing ability, while keyboarding skills are not pertinent to the measurement of writing.

The second approach, forced acceptance, acknowledges the role of technology to a limited extent by implicitly incorporating basic computer literacy skills (e.g., reading on a computer screen and the use of keyboarding skills) into the constructs. This approach is primarily motivated by the need to use computer-based delivery to scale up testing and is commonly adopted in large-scale testing, although test designers remain concerned about the potential impact of technology skills on test performance.

The third approach, cautious acceptance, may allow controlled use of assistive tools by test takers, such as spell checkers, grammar checkers, and online dictionaries. However, this approach has not yet been implemented in large-scale testing, except that online glossaries for pre-selected words may be made available, for example, in the Progress in International Reading Literacy Study (PIRLS) assessment.

The final approach, progressive embracing, takes the most advanced stance by defining the constructs as communication skills fully integrated with computer and digital information literacy skills. This includes the ability to effectively use a full set of editing and generative AI tools to accomplish communication tasks, as well as the ability to use digital technologies to search for, identify, evaluate, organize, refine, and synthesize information to fulfil a task.

Expanded constructs in relation to AI technology

Building on the fourth approach discussed earlier, it appears that, despite advocacy from researchers, its practical application in large-scale testing contexts still seems far-fetched. While learners are increasingly using generative AI in their daily communication, these tools are often initially labeled as potential cheating aids in assessments. However, a more nuanced perspective suggests that integrating these tools into assessments in contexts where they are used to enhance communication could reflect more accurately their growing prevalence in everyday communication.

Language tests aim to closely replicate real-world communication. The widespread adoption of assistive and generative AI tools in everyday communication has stimulated fresh investigations about how communicative competence should be defined and prompted language testers to reconsider the nature of language competence and how testing practices can be transformed. The growing use of assistive and generative AI tools in real-life communication has sparked debates and research inquiries into the nature of communicative competence, as highlighted in the special issue of "Advancing language assessment with AI and ML" recently published in *Language Assessment Quarterly* (Voss et al., 2023; Xi, 2023).

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

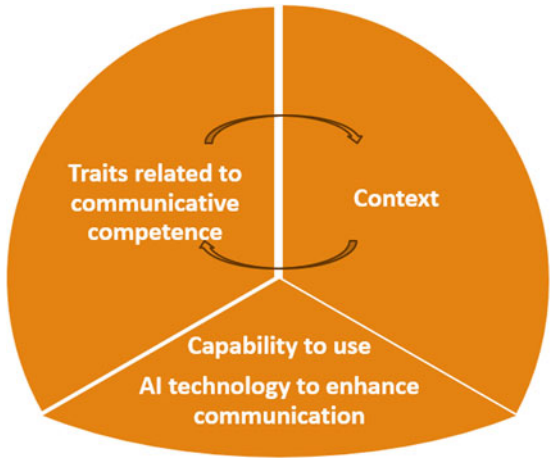

Figure 1. AI-mediated interactionalist approach.

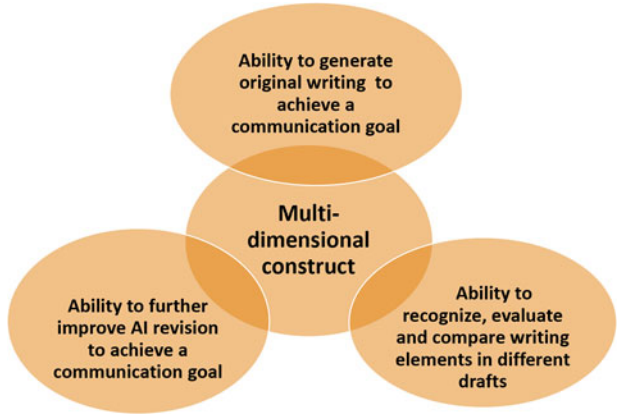

Figure 2. The multidimensional construct of AI-mediated communicative competence.

Earlier the interactionalist approach to construct definition was discussed, which emphasizes the interplay between traits and context. Integrating AI technology into this framework broadens the definition of communicative competence to include proficiency in utilizing AI tools and interpreting and utilizing AI-generated outputs and feedback to enhance communication (Figure 1). Given AI's capability to generate sophisticated content and provide comprehensive feedback, this expanded definition includes not only the effective use of AI tools to produce intended outputs but also, more critically, the ability to analyze, compare, evaluate, and modify these outputs to improve communication.

To further elucidate the capability to use AI technology to enhance communication, it is essential to recognize that this capability encompasses both AI literacy skills and broadened linguistic and cognitive skills (Figure 2). AI literacy skills involve understanding how AI tools work, prompting AI tools to obtain desired outputs, and refining

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

those outputs through re-prompting. These skills require cognitive processes to formulate clear and appropriate commands, evaluate initial AI outputs, identify areas that do not meet expectations, and determine whether further prompting is necessary. Additionally, AI literacy skills include navigating AI tool environments with an adequate understanding of potential ethical issues, being aware of the implications of using AI, ensuring data privacy, and understanding the biases that may exist in AI algorithms.

Broadened linguistic and cognitive skills are also required to interpret and refine AI-generated outputs. These skills, depending on how AI tools are being used, may involve, for example, analyzing AI-generated content, comparing it to the intended message, and making necessary adjustments to ensure clarity, coherence, and appropriateness of the message. In other words, effective use of AI tools in communication may entail the ability to critically evaluate the quality of AI outputs, and then adapt and incorporate them meaningfully into one's own work.

While some might argue that AI literacy skills should not be conflated with linguistic skills, the widespread use of AI tools by language users today suggests that the traditional notion of communicative language ability may no longer be entirely adequate. As the use of AI tools becomes an integral part of daily communication, proficiency in utilizing these tools and understanding and using their outputs is crucial for a comprehensive re-definition of communicative competence. This expanded definition provides guidance for reformulating language tests to measure these skills alongside conventional ones and to support the intended use of the tests.

Take language tests used for higher education admissions as an example. Nowadays, university students have access to editing and generative tools to refine or even generate a substantial basis for their language outputs, such as class presentations and course papers. Admittedly, higher education institutions across the globe may have varying policies regarding the use of generative AI tools by students in writing their course papers for content courses. While AI tools for stylistic and content editing to enhance clarity and coherence are generally permitted, plagiarized ideas and content are strictly prohibited. Consequently, continuing to focus tests solely on conventional communication skills would not help to select those students most likely to succeed in coping with real-world communication demands at a university. Instead, next-generation language tests should evaluate how well students can use these AI tools to enhance their language production, reflect on and refine AI-generated content, and effectively communicate in diverse academic contexts.

Example scenarios of using AI tools and corresponding changes in constructs

AI tools can be used in many ways to enhance communication, each leading to somewhat different changes to the constructs. One possible scenario involving the use of AI is having learners draft an initial version of their essay, submit it, and then utilize generative tools to enhance language, content, and audience awareness (see Scenario 1). After this, learners would assess and compare the original draft with the AI revision, make further refinements, and finally submit the completed composition. This is similar to a setup in a dynamic assessment (Poehner & Lantolf, 2013, 2021), where learners' original outputs as well as updated ones with the help of AI tools are

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

collected for the assessor to not only examine the differences but also evaluate how each learner is able to reflect on and evaluate AI's revisions to ultimately improve their writing.

Another scenario involves using AI to generate multiple outputs on the same topic tailored for various audiences. For example, learners might use AI to generate a letter to the principal arguing against mandatory school uniforms and another article on the same topic for the school newspaper. Learners would then compare the two pieces, identifying differences in language use, rhetorical devices, and tone and voice appropriate for each target audience. Then, they would ask AI to analyze the differences and identify potential improvements in the two AI-generated outputs. Following this, they would compose a short post presenting their stance on an online student discussion forum. This exercise would help learners compare differences in wording, formatting, structure, content emphasis, and tone and voice.

Scenario 1:

(a) Ask user to write for a purpose and an audience and submit their original writing
(b) Ask AI to generate feedback on language, content, and audience awareness
(c) Revise and resubmit the writing

Scenario 2:

(a) Ask AI to generate writing outputs targeting two audiences
(b) Ask user to compare the two outputs
(c) Elicit AI feedback on the differences
(d) Ask the user to evaluate AI generated outputs and analysis of feedback on the differences and identify potential areas of improvements in the two outputs
(e) Ask the user to adapt the writing for a different audience

The additional language communication skills required in Scenario 1 extend beyond creating original writing to achieve a communication goal. When working with AI-generated revision or feedback on users' original writing, these skills include the ability to compare, analyze, and interpret AI-generated outputs, assess whether these outputs represent improvements or deviations from the writer's original message, and determine whether they enhance or deteriorate the quality of the outputs. This involves making informed decisions about accepting, rejecting, or further revising the outputs. When comparing two pieces of AI-generated writing in Scenario 2, cognitive functions involved include analyzing, evaluating and comparing linguistic and other characteristics to facilitate task completion. Comparing the writing to two audiences can increase awareness of the differences in relation to varying purposes and audiences and provide a model for adapting it for a different audience.

The previously discussed application scenarios indeed warrant a multidimensional conception of writing ability. This encompasses traditional aspects of writing proficiency, such as adherence to writing conventions, structural coherence, the quality of arguments, and the ability to adapt the content, style, voice, and tone of writing for different audiences. Moreover, it includes a new dimension, such as the ability to identify,

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

evaluate, and compare different pieces of writing, and to analyze differences in writing quality. This process involves analytical and evaluative skills as well as editing skills.

Furthermore, a third dimension is introduced, which entails the ability to edit language and content created by AI tools, in order to complete the writing task. This comprehensive approach ensures a holistic assessment of writing abilities, integrating both traditional and expanded skills as a result of using AI tools in the process of their language production.

These scenarios illustrate how AI can be integrated into language assessments to reflect demands of contemporary communication, fostering and measuring audience-specific communication. Such language test tasks also prepare students for real-world communication, where the ability to tailor messages to various audiences and refine content using AI tools is increasingly valued.

A nuanced framework for integrating the use of AI tools into assessments

In discussing the use of editing and generative tools by test takers, it is essential to adopt a nuanced and granular framework due to the numerous factors that need consideration. These factors include the choice between editing and generative tools, the stakes of the test (high-stakes versus low-stakes), and the proficiency level of the target test takers (beginners, intermediate, or advanced). Taking these factors into account allows us to rethink the proficiency levels of the students, design new types of assessments and criteria, and reconsider the interpretation and use of scores.

Recent investigations in the field of second/foreign language research indicate that low-level students may struggle with the sophisticated language generated by AI, finding it challenging to make meaningful and productive use of AI tools to enhance their own writing (Woo et al., 2025). This realization calls for differentiated approaches to assessments targeting different proficiency levels, as communicative competence is redefined and assessment practices are transformed. The construct mediated by AI technology becomes fluid when applied to assessments designed for varying proficiency levels.

In Table 1, I have proposed whether candidates should be provided with editing or generative tools, depending on the test purpose and target level of the test (Xi, 2023). For assessments aimed at low-proficiency test takers, the integration of generative AI tools may be overwhelming. For beginners in a low-stakes assessment, it is advisable to provide editing rather than generative tools. This is because assessments of beginner-level writing often emphasize conventions of writing, such as capitalization, punctuation, and grammar, which cannot be effectively assessed if test takers are allowed to use generative tools. While the evaluation may well include overall communicative effectiveness such as task completion depending on the purpose of the assessment, a heavy focus is placed on these conventions. Therefore, limited use of editing or polishing tools, captured and monitored during the assessment process, can be considered to expand the construct. Correspondingly, the instructional focus should prioritize developing foundational skills, particularly the conventions of language production, to better prepare learners for tasks involving more sophisticated use of both editing and generative tools.

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

Table 1. Access to AI tools by test takers in combinations of target level and test stakes (adapted from Xi, 2023)

Test stakes	Target level	Access to editing tools	Access to generative tools	Notes
Low stakes	Low	Yes	No	Focus on assessment of conventions
	Intermediate	Yes	Yes	With/without assistance for dynamic assessment
	Advanced	Yes	Yes	
High stakes	Low	No	No	Focus on assessment of conventions
	Intermediate	Yes	No	Redefine constructs Revamp rubrics Reformulate interpretation & use
	Advanced	Yes	Yes	

For tests targeting intermediate-level test takers, who possess the basic skills to use generative tools and work with well-edited AI-generated outputs tailored to their levels, the emphasis could shift to nurturing their ability to use generative tools effectively and process AI outputs that match their proficiency levels.

Conversely, for tests designed for upper intermediate and advanced test takers, full integration of generative tools into the test construct can be considered. The instructional emphasis should be on enhancing their ability to use generative tools effectively and perform additional tasks based on AI-generated outputs to meet specific communication needs.

In all these designs that allow access to editing and/or generative tools, a dynamic assessment approach can be employed. This approach involves students initially writing without any tools and then revising with tools. By comparing the writing samples, we can gain insights into students' current proficiency levels and their potential for future development. This method provides a comprehensive evaluation of their ability to utilize AI tools and their overall communicative competence.

Such a dynamic approach not only helps in assessing the students' current language skills but also prepares them for real-world scenarios where they will likely use AI tools to enhance their communication. By integrating these tools into the assessment process, we can provide a more relevant measure of their true communicative competence and better equip them for future communication demands.

Barriers to the conceptualization and operationalization of the expanded construct and a path forward

Despite conceptual advances in redefining communicative competence in today's digital era (Voss et al., 2023; Xi, 2023), the integration of such advancements into large-scale testing, especially tests that target learners with diverse backgrounds around

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

the world, will remain a prolonged process. Several barriers impede the conceptualization and operationalization of approaches that incorporate AI technology into communicative competence, including the previously discussed Cautious Acceptance and Progressive Embracing approaches.

Conceptually, longstanding views on language constructs are deeply entrenched. Traditional conventions, such as writing norms, remain integral to writing curricula globally and are central to writing assessments designed for learners at varying levels of writing proficiency. For example, the Score Level 3 (on a 0–5-point scale) descriptor of the TOEFL iBT Independent Writing Task Rubric includes "some noticeable lexical and grammatical errors in sentence structure, word form, or use of idiomatic language" (Educational Testing Service, n.d.). Similarly, the Band 6 descriptor of the essay writing task of the IELTS test evaluates both lexical resource and grammatical range and accuracy, noting that "There are some errors in spelling and/or word formation, but these do not impede communication" and "Errors in grammar and punctuation occur, but rarely impede communication" (British Council, May 2023). These scoring rubrics highlight that skills related to writing conventions are fundamental to writing proficiency. The use of assistive tools such as grammar and spell checkers is often perceived by mainstream language testers as potentially compromising or complicating the assessment of language skills. Generative tools, such as ChatGPT, are frequently viewed as cheating aids, especially in tests that have consequences, such as those that influence course grades or admissions. In writing composition classes focused on developing writing skills, allowing students to use generative AI tools in an unproctored or unmonitored environment is controversial (Perkins, 2023). As a result, such tools are prohibited in many classroom assignments and assessments and are entirely off-limits in large-scale tests to ensure that test-takers' work genuinely reflects their individual knowledge, skills, and abilities.

Moreover, disparities in access to technology and unequal opportunities to practice with AI tools raise significant fairness concerns. Learners from underprivileged backgrounds often lack access to advanced technology, limiting their ability to familiarize themselves with AI tools that facilitate communication. This digital divide can exacerbate existing educational inequalities, widening the gap between underprivileged and privileged learners, as those with better access to technology can leverage these tools to enhance their learning, and consequently, their assessment performance. These disparities pose major challenges to innovation in large-scale testing. Performance on assessments may be impacted by unequal opportunities to learn and might reflect test takers' limited ability to use AI tools effectively to enhance their communication, rather than their independent language abilities, when their original language outputs are not elicited and captured in an assessment.

In practice, the current competitive landscape of international English language testing, which has been most dominant and influential in the L2 testing field, fuels a race to create the most user-friendly tests to attract test takers (Educational Testing Service, January 2024; Pearson, n.d.). Integrating AI as facilitative tools for test-takers in testing requires substantial investment to customize AI technology for language tests and requires a comprehensive research program to adequately support the validity, fairness, and ethical use of the technology, which currently may not be a priority for

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

large-scale English testing programs. While AI can enhance the efficiency and potentially the quality of test development, administration, scoring, and feedback provision, helping test providers manage costs, streamline operations, and position themselves as innovators in AI technology, the extensive research and development necessary for the comprehensive integration of AI tools into the test-taking experience requires significant increases in development and operational budgets. Additionally, it introduces complications in test design, scoring, score interpretation, and use as well as controversies and challenges regarding the potential validity and fairness of a test, which leads to it being deprioritized in current large-scale English language testing practices.

Currently, there is no established precedent for allowing students to use editing tools in large-scale language tests. Our existing writing assessments place excessive emphasis on writing norms and conventions. Automated scoring tools, which often rely on the high correlation between lower-order and higher-order skills, achieve high accuracy in predicting human scores by focusing on features associated with lower-order skills such as length, grammar, and vocabulary (Chen et al., 2018) and surface-level features as proxies for higher-order skills such as organization and development (Chen et al., 2016). However, automated scoring engines for tests that permit test takers to use AI tools would need to assess expanded AI literacy, linguistic, and cognitive skills, presenting significant challenges. The prevalent design of writing assessments, combined with the current state of automated scoring tools, has perpetuated the tendency to maintain the status quo in language testing.

Providing learners and test-takers with access to AI tools represents a significant leap of faith for many language educators and testers. The rapid advancement of AI technology has outpaced the evolution of language constructs and testing practices, creating discomfort and uncertainty among some educators. This rapid pace of technological development has led to a gap between the potential of AI tools and their integration into language assessment frameworks. A recent survey found that 36% of L2 teachers were not aware of any institutional policies for responsible and ethical use of AI as well as appropriate safeguards (Galaczi & Luckin, 2024). Without the guidance of established test theories and guidelines specifically addressing the use of AI, classroom assessment designers may have to rely on their professional judgment to navigate this uncharted territory.

Amid ongoing debates about the use of AI tools by test takers in assessments, some more forward-looking perspectives have emerged. Oppenheimer (2023) suggests that rather than focusing on concerns about student cheating with AI tools, we should address the issue by restructuring assessments and classes so that using these tools is not deemed dishonest. He also emphasizes that since generative AI tools are increasingly integrated into students' daily lives and professional environments, assessments should be designed to evaluate how individuals solve problems utilizing tools like ChatGPT (Oppenheimer, 2023).

In a similar vein, in the language learning and assessment field, Galaczi and Luckin (2024) argue that generative AI enables a shift from the conventional "learn, pause, test" model to a more integrated approach where learning and assessment are interconnected, leveraging the unique affordances of generative AI and rooted in a communicative language methodology.

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

Navigating the current landscape of language assessments requires a high degree of adaptability and innovation from educators and language testers. Classroom teachers, in particular, have demonstrated considerable flexibility in incorporating AI tools into their assessments. They actively engage with these tools to enhance the learning and assessment process while attempting to navigate the many ambiguities and challenges associated with their use. These challenges include determining the appropriate, responsible, and ethical use of AI tools, ensuring that their integration does not compromise the integrity of the assessment, and maintaining fairness and validity in assessment.

The integration of AI tools in classroom assessments has the potential to revolutionize the way language skills are taught and evaluated. By leveraging AI, educators can create more dynamic and interactive assessment tasks, leading to a more contemporary approach to assessment.

However, the path to fully integrating AI into language assessments is anticipated to be fraught with challenges. As argued earlier, a key challenge is that disparities in learners' digital access and AI literacy levels introduce significant concerns regarding test validity and fairness, particularly in high-stakes, large-scale testing. Addressing these disparities requires deliberate efforts in educational policy and testing reform as well as changes in testing practices. Efforts may include:

- developing guidelines and best practices for integrating generative AI tools into assessments
- introducing initial innovations to large-scale testing in local contexts involving homogeneous populations
- standardizing, monitoring and logging the use of AI tools in large-scale testing
- promoting more equitable access to AI technology and AI literacy around the world.

First, the absence of comprehensive guidelines and established best practices means that educators must navigate this complex landscape on their own initially. This can lead to inconsistencies in how AI tools are used in assessments and how communicative competence is assessed, potentially affecting the validity, reliability, and fairness of assessments. Additionally, the ethical implications of using AI in educational settings must be carefully considered, particularly in terms of data privacy, the potential for bias in AI algorithms, and the impact on student engagement during assessment.

So far, several testing associations and test providers have published guidelines for responsible and ethical use of AI in testing (e.g. Association of Test Publishers, 2022; Burstein, 2023; Educational Testing Service, n.d.b; Galaczi & Luckin, 2024; Xu et al., 2024). However, none of these standards and guidelines are geared toward classroom assessment, providing limited guidance to classroom teachers on how to use AI in classroom assessments. Additionally, they do not address the need to rethink the concept of communicative competence or other skills due to the use of AI tools by test takers, with the exception of Galaczi and Luckin (2024), who raise the need for designing language assessments that incorporate the use of generative AI tools. Thus, no practical guidance is provided to teachers about what tools to

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

use in assessments based on the purpose of assessment, and how the use of tools by test takers can be integrated into assessment task design.

While the integration of AI tools into language assessments holds great promise, it requires careful consideration and a balanced approach. Educators and testers must work together to develop new frameworks and guidelines that address the unique challenges and opportunities presented by AI technology. By doing so, they can ensure that AI tools enhance rather than undermine the assessment process, providing learners with more meaningful experiences that prepare them for the demands of the digital age. This is one of the first areas where language testing theorists need to collaborate closely with teachers and practitioners to provide guidelines and best practices for incorporating generative AI tools into assessments. These guidelines and best practices, initially geared toward classroom assessment, can then be updated for high-stakes, large-scale testing as the test purposes, practicality considerations, and consequences are different.

Reforms to our large-scale language testing practices must continue due to their catalytic and systemic impact on local teaching and learning practices. The impetus for these reforms will stem from a combination of advances in test theories and transformations in classroom assessment practices. Theoretical advancements will offer increasing clarity and guidance on the conceptual challenges discussed above. Simultaneously, evolving classroom assessment practices will exert pressure on large-scale testing, promoting eventual changes. One possible pathway is to tackle this challenge initially within local large-scale testing where the technological infrastructure is more advanced, and the learner population is relatively homogeneous. Such testing environments, including, for example, tests for admissions or workforce qualifications, in regions where learners more frequently utilize AI tools, can begin to reform testing practices relatively quickly. In contrast, global language tests, taken by learners with varying levels of access to and familiarity with generative AI tools, are likely to take much longer to adapt. However, this evolution is not improbable – much like how computer-based language testing has now become standard in large-scale global English assessments, even though paper-based testing was dominant just two decades ago.

A third area to address is to tackle some of the operational challenges associated with the diverse use of generative AI tools by test takers and the inferences we can make about their independent and aided ability. As discussed earlier, a dynamic assessment design can allow evaluators to see both learners' original writing and revised writing based on AI generated feedback/outputs. To achieve standardization in large-scale testing, it is important to provide test takers with the same generative tool(s) and the same set of AI prompting language to begin with, ensuring consistent administration conditions across test takers. With the availability of test-taking process monitoring software technology, it is feasible to monitor the entire composition process by test takers, including identifying which parts are original writing, which parts are AI-generated, and which are revisions based on AI outputs. These consistent administration conditions as well as the ability of the testing system to monitor and log the use of AI tools by test takers will help to disambiguate some potential confusion while making inferences about learners' communicative competence.

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

Ultimately, addressing the digital divide across the globe is essential to mitigating current inequities in digital access, which would significantly impact the fairness of a language test with generative AI tools incorporated. To support global learners, it is imperative to improve technological infrastructure and to develop and implement local digital education programs aimed at advancing AI literacy, especially for marginalized groups (Gonzales, 2024). This calls for close collaboration among governments, educational institutions, local communities, and technology companies (Gonzales, 2024). In the language testing field, resources should be devoted to providing these AI literacy training programs to teachers, who can subsequently train their students, helping to promote digital inclusion for all learners globally.

While classroom assessments largely rely on off-the-shelf AI tools, large-scale testing providers are in a unique position to develop customized solutions. If these providers can realign their strategic priorities and allocate financial and human resources accordingly, they have the potential to innovate and lead in the field of assessment. Those large-scale providers that act promptly will be well-positioned to reclaim important ground in driving assessment innovations.

Ultimately, the combination of theoretical advancements and practical classroom transformations is expected to catalyze significant changes in large-scale language testing. This synergy between theory and practice will help ensure that assessment practices remain relevant and effective in measuring communicative competence in a rapidly evolving digital landscape. By embracing these changes, large-scale testing providers can not only enhance the validity and fairness of assessments but also better support the educational community in preparing learners for the demands of the modern world.

Conclusion

The perception of technology's role in language assessment has evolved significantly over time. Initially seen as a source of construct-irrelevant variance, technology was believed to potentially contaminate the assessment of language skills. However, as technology advances, particularly with the advent of AI and machine learning, its potential to enhance language assessment has become increasingly apparent. Generative AI tools have demonstrated their ability to produce sophisticated text and speech and provide meaningful feedback to enhance real-life communication, challenging traditional notions of language competence.

In the last few years, real-world communication has changed dramatically, particularly with the rise of AI technology, which many now use to enhance or generate significant parts of their communication. This shift has led researchers to reconsider the nature of communicative competence in today's digital age, prompting a reassessment of what language tests should evaluate and how communicative competence should be measured.

Specifically, integrating AI tools into language assessments can help mirror real-world practices and assess higher-order skills currently overlooked, such as analytical and evaluative skills in writing. As AI advances to enhance asynchronous or real-time oral communication via tools like prepared speech coaching, instant hints for real-time interaction, the definition of oral communicative competence will also evolve, enabling

https://doi.org/10.1017/S0267190525000078 Published online by Cambridge University Press

language testers to measure currently unassessed skills. This approach not only aligns with technological advancements but also provides a more thorough assessment of communicative competence.

Providing learners and test-takers with AI tools represents a significant step for many language educators and testers. The swift advancement of AI technology has surpassed the development of language constructs and testing practices, causing discomfort among some educators. In the absence of established test theories, classroom assessment designers have had to rely on their professional instincts to integrate AI tools into language assessments. Classroom teachers have shown greater adaptability, adopting these AI tools in their assessments while grappling with the complexities surrounding their appropriate, responsible, and ethical use.

The redefinition of communicative competence has significant implications for language assessment. Innovation in high-stakes exams tends to progress slowly due to various challenges. Low-stakes formative assessments provide a fertile ground for experimentation and innovation. In these contexts, educators can explore the use of various AI tools, such as editorial and generative tools, to enhance assessment. For instance, editorial tools can help demonstrate students' ability to use them to improve the efficiency and accuracy of their writing, while generative tools, if incorporated effectively in assessments, can help measure students' AI literacy skills as well as expanded linguistic and cognitive skills to interpret and use AI outputs to improve the quality of their language output.

Collaboration among language researchers is crucial to keeping assessments in line with real-world communication developments. Unlike the traditional method of driving curriculum changes through large-scale test innovations, it is likely that formative assessment designers will lead this digital transformation. By incorporating AI tools into formative assessments, they can pave the way for future changes in large-scale language tests.

References

Alderson, C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, *13*(3), 280–297. https://doi.org/10.1177/026553229601300304.

Association of Test Publishers. (January 2022). *Artificial intelligence principles*. Retrieved March 30, 2025, from https://www.testpublishers.org/ai-principles.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche & D. Bayliss (Eds.), *What are we measuring? Language testing reconsidered* (pp. 41–71). University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

British Council. (May 2023). *IELTS writing band descriptors*. Retrieved March 30, 2025, from https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf.

Burstein, J. (2023). *The Duolingo English Test Responsible AI Standards*. Retrieved March 29, 2024, from https://go.duolingo.com/ResponsibleAI.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp.30–40). Center for Applied Linguistics.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369–383. https://doi.org/10.1191/0265532203lt264oa.

Chan, S., Bax, S., & Weir, C. (2017). Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. *IELTS Research Reports Online Series*, *4*, 1–47.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman, and A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge University Press.

Chen, J., Fife, J. H., Bejar, I., & Rupp, A. A. (2016). Building e-rater® scoring models using machine learning methods. *ETS Research Report Series*, 2016.

Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., … Binod, G. (2018). *Automated scoring of nonnative speech using the SpeechRaterSM v. 5 engine*. Princeton: Educational Testing Service.

Educational Testing Service. (January 2024). Exciting enhancements to the TOEFL iBT® test. https://www.etsglobal.org/mc/en/blog/news/toefl-ibt-enhancements.

Educational Testing Service. (n.d.a). *TOEFL iBT® writing for an academic discussion rubric*. Retrieved March 30, 2025, from https://www.ets.org/pdfs/toefl/toefl-ibt-writing-rubrics.pdf.

Educational Testing Service. (n.d.b). *Responsible use of AI in assessment: ETS principles*. Retrieved March 30, 2025, from https://www-vantage-stg-publish.ets.org/Rebrand/pdf/ETS_Convening_executive_summary_for_the_AI_Guidelines.pdf.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Figueiredo, S. (2021). Second language testing: A psycholinguistic approach. *International Journal of Childhood Education*, *2*(2), 26–31. https://doi.org/10.33422/ijce.v2i2.15.

Galaczi, E., & Luckin, R. (2024). *Generative AI and language education: Opportunities, challenges and the need for critical perspectives*. Cambridge Papers in English Language Education.

Gonzales, S. (2024, August). *AI literacy and the new digital divide - A global call for action*. Paris, France: UNESCO. Retrieved March 30, 2025, from https://www.unesco.org/en/articles/ai-literacy-and-new-digital-divide-global-call-action.

Green, A. (2003). *Test impact and English for academic purposes: A comparative study in backwash between IELTS preparation and university pre-sessional courses* [Unpublished PhD thesis, University of Surrey]. Centre for Research in Testing, Evaluation and Curriculum in ELT, University of Surrey.

Green, T., & Maycock, L. (2004). Computer-based IELTS and paper-based versions of IELTS. *Research Notes*, *18*, 3–6.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249. https://doi.org/10.1080/15434303.2011.565844.

Jin, Y. (2000). The washback effects of College English Test-Spoken English Test on teaching. *Foreign Language World*, *118*(2), 56–61.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. Longman.

Leung, C. (2022). Language proficiency: From description to prescription and back? *Educational Linguistics*, *1*(1), 56–81. https://doi.org/10.1515/eduling-2021-0006.

Long, M. (2015). *Second language acquisition and task-based language teaching*. John Wiley & Sons.

Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230–244. https://doi.org/10.1017/S0267190516000027.

Oppenheimer, D. (January 2023). ChatGPT has arrived – And nothing has changed. *Times Higher Education*. Retrieved March 30, 2025, from https://www.timeshighereducation.com/campus/chatgpt-has-arrived-and-nothing-has-changed.

Pearson. (n.d.). *PTE academic just got better!* Retrieved March 30, 2025, from https://www.pearsonpte.com/articles/pte-academic-just-got-better.

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, *20*(2). https://doi.org/10.53761/1.20.02.07.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research*, *17*(3), 323–342. https://doi.org/10.1177/1362168813482935.

Poehner, M. E., & Lantolf, J. P. (2021). The ZPD, second language learning, and the Transposition ~ Transformation dialectic. *Cultural-Historical Psychology*, *17*(3), 31–41. https://doi.org/10.17759/chp.2021170306.

Read, J., & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. In R. Tulloh (Ed.), *International English Language Testing System (IELTS) research reports* (Vol. 4, pp. 153–206). IELTS Australia

Saville, N., & Hawkey, R. (2004). The IELTS impact study: Investigating washback on teaching materials. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 73–96). Lawrence Erlbaum Associates.

Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications*, 1–40. https://doi.org/10.1007/s11042-024-20016-1.

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

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. ETS Research Report Series, i–30. https://doi.org/10.1002/j.2333-8504.1998.tb01757.x.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, *49*(2), 219–274. https://doi.org/10.1111/0023-8333.00088.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344. https://doi.org/10.1177/0265532211424478.

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.

Wall, D., & Horak, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study*. TOEFL Monograph Series, MS-34. Educational Testing Service.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, *13*(3), 318–333. https://doi.org/10.1177/026553229601300306.

Woo, D. J., Susanto, H., Yeung, C. H., Guo, K., & Huang, Y. (2025). Approaching the limits to EFL writing enhancement with AI-generated text and diverse learners. *arXiv preprint* arXiv:2503.00367. https://doi.org/10.48550/arXiv.2503.00367.

Xi, X. (2015, March). Language constructs revisited for practical test design, development and validation. Paper presented at the 37th Language Testing Research Colloquium, Toronto, Canada.

Xi, X. (2023). Advancing language assessment with AI and ML–leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, *20*(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488.

Xi, X., Bridgeman, B., & Wendler, C. (2024). Evolution and future trends in tests of English for university admissions. In A. J. Kunnan (Ed.), *The concise companion to language assessment* (pp. 355–370). Wiley-Blackwell.

Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. E. (2021). Assessing academic speaking. In X. Xi, and J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (1st edn, pp. 152–199). Routledge. https://doi.org/10.4324/9781351142403.

Xu, J., Schmidt, E., Galaczi, E., & Somers, A. (2024). Automarking in language assessment: Key considerations for best practice. In *Cambridge papers in English language education*. Cambridge University Press & Assessment.

Young, R. F. (2008). *Language and interaction: An advanced resource book*. Routledge.

Cite this article: Xi, X. (2025). Revisiting communicative competence in the age of AI: Implications for large-scale testing. *Annual Review of Applied Linguistics*, *45*, 200–221. https://doi.org/10.1017/S0267190525000078