

# **AI in Large-Scale International Language Testing: Challenges and Opportunities**

**Michael Milanovic**

**Plenary delivered at  
The Eighth Conference on  
English as a Foreign Language Teaching  
and Assessment:  
School of Foreign Language and  
Literature, Guiyang, China  
July 25-26, 2025**


## Background

In this paper, I'm going to be exploring the impact of AI on language assessment. My focus will be on international exams for study, employment and migration although many of my observations may be more generally relevant. For the most part, I'll be talking about AI systems that affect language use such as ChatGPT, Deepseek, Claude etc.

The language testing landscape is changing continuously and now is a particularly interesting time as we see how AI is shaping up to have a big impact. Back in 1982 when I first came to China we were in a paper-based, multiple-choice, classical test analysis world, a tradition developed in the 1950s and still alive and well today in many contexts. Communicative language testing was just starting to have an impact and ELTS, developed in the 1970s (the mother of IELTS), became the best-known communicative language test. In the 1980s, while TOEFL was testing hundreds of thousands of candidates a year worldwide, ELTS was testing around ten thousand a year and in its 15-year life only two test forms were ever used. Researchers were looking at construct validity (the extent to which a test is measuring the language skills it is supposed to measure), worrying about the unitary competence hypothesis and focusing on test reliability.

In the 1990s, everyone was modernising, developing and improving item banking systems, calibrating test items and test forms using Item Response Theory (IRT). People were looking at how tests could be delivered by computer but generally in a local PC context as the internet was so unstable and restricted. In 1997, TOEFL launched its first CB test. In the international context, high stakes testing for study and migration dominated the language testing world. Test delivery was starting to become a global industry particularly in China and India, reflecting the rapid growth of these economies and many others around the world. From an academic perspective, the 1990s represented a period of consolidation and theoretical development in language testing, moving away from largely psychometric approaches towards more holistic, communicative, and socially conscious testing practices. Many contemporary concerns in language assessment today – particularly around validity, authenticity, and the social impact of testing – were developed in the 1990s.

In China, the 1990s were a time of modernisation of English language testing methods. A greater focus on communicative language ability became a priority




with the development and introduction of the Public English Language Testing System (PETS). This was a joint project funded by the governments of China and the UK between 1997 and 2000. At the time there were 11 English language tests in general use in China and the PETS five level system, drawing heavily on the communicative language testing movement and the newly developed Common European Framework of Reference (CEFR). PETS was intended to replace this set of tests with a single five-level system; I think, however, that we ended up with 16 tests as opposed to 11. I was quite heavily involved in the development of the CEFR as well as the PETS tests.

The next decade saw a massive increase in international assessment both in national education systems, where English became the most widely taught language in the world and most used in international study and migration. Whereas, in 2000, less than a million people were taking IELTS and TOEFL combined, this number rose to around five million by 2010. Much greater focus turned to managing the rapidly increasing test candidature, to test security, cheating, item harvesting, cramming and the impact this had on test validity and integrity. TOEFL edged towards fully automated CB assessment while IELTS continued to focus on authenticity and positive impact and remained paper-based for the most part. It is not clear how much this difference in focus made to test takers but during this decade we saw IELTS overtake TOEFL and become the most widely used test for international study and migration. Technology played a role in test delivery and test security. An excellent example of this is China's massive network of CCTV monitoring of high stakes tests.

In the background, from an academic perspective, testing was starting to migrate to computer-based delivery. Most existing pencil and paper test tasks were adapted for computer-based use and there was a focus on the automated marking of writing and speaking. The language testing field continued to professionalise with a particular focus on assessment literacy and standards. Frameworks such as the CEFR were introduced and there was a focus on validity and validation research that continued to build on Messick's (1989) validity framework and Kane's (2013) argument-based approaches.

The first major disruption in the decade came with the introduction of the Pearson Test of English (PTE) in 2009. PTE was the first fully automated test to be used in the context of international study and migration. Its innovative impact was not really on test materials, which remained somewhat traditional, if not rather old fashioned, but rather on the fact that for the first time, the human




being was removed entirely from the equation when it came to marking across all four skills. It is not clear that the test was very well received amongst teachers and students and in terms of candidature, for most of the next decade, it remained a distant third behind IELTS and TOEFL. However, it became increasingly popular in countries such as Australia and India as we approached 2020 because of its availability and accessibility. While its delivery remained anchored in the traditional high stakes test centre network context, it could be made available several times a day in places where demand was high.

In the 2020s, there have been significant developments, and the pace of change is accelerating. Firstly, the concept of testing at home and being remotely proctored was widely adopted in several contexts. This was driven largely by the pandemic. LanguageCert pioneered online proctored tests (OLP), taken by thousands worldwide, including in China. Secondly, we have seen AI grow from relatively simple operations, largely inconsequential to language testing, to systems which challenge the very fabric of language assessment.

International testing organisations collectively administer millions of high-stakes international language tests annually across diverse global contexts - high stakes tests that determine university admissions, immigration opportunities, and professional certification, and there is a growing pressure to offer these tests in an online proctored (OLP) environment. LanguageCert was the first organisation to run secure high stakes OLP language tests in 2019, and you just have to look at the subsequent introduction of remotely proctored online tests by Pearson (Express), ETS (Home), IDP (Envoy), the British Council (Aptis Remote), and others to see how much of a growing trend OLP is.

AI technologies have the potential for negative impact as they present significant challenges to testing systems, while at the same time also offering innovation opportunities. They represent a technological disruption that goes beyond previous innovations in computer-based testing. Such innovations were relatively low key and revolved largely around delivery. Item types in listening, reading and writing didn't really change much and when it came to speaking, assessments focused on pronunciation, repetition, read aloud and so on rather than the communicative use of the language. If anything, use of technology has typically played a restrictive role in testing because it couldn't do anything very creative.

AI models on the other hand, are quite different because they can generate human-like text, they can answer comprehension questions with high accuracy, create barely identifiable deep fakes in oral interview contexts and even simulate



conversational exchanges. I'm going to look at how AI might impact large-scale international assessments, paying attention to validity, fairness, and security. I'm also going to touch on some potentially relevant adaptation strategies.

## Validity Considerations


Large-scale international language testing organisations are required to show that their scores accurately reflect relevant language abilities and support appropriate decision-making by universities, employers, and immigration authorities worldwide. In other words, they need to demonstrate that the tests are measuring what they are supposed to measure: that they have construct validity.

As AI systems like ChatGPT, Deepseek and others become increasingly prevalent, they are starting to change the communicative competence construct. The workplace integration of AI tools is accelerating rapidly: according to Microsoft's (2024) *Work Trend Index* (taken by 31,000 people across 31 countries), 75% of knowledge workers now use AI at work, with 46% having adopted it within just the past six months. 90% of users report that AI helps them save time, 85% that it helps them focus on their most important work, and 84% that it helps them be more creative.

Similarly, PwC's (2024) *Global Workforce Hopes & Fears Survey* (taken by 56,600 individuals across 50 countries and territories), found that more than 70% of generative AI users agree that AI tools create opportunities to be more creative at work and improve the quality of their work. These findings indicate that AI-mediated communication is becoming normalised in professional contexts, suggesting that traditional constructs will need to evolve as the target language use domain changes fundamentally. This represents a qualitatively different shift from previous technological advances. Electric typewriters may have increased speed and word processors improved spelling accuracy, but AI is fundamentally changing how we work with and manipulate knowledge and language.

What does it mean in the assessment context? Will the assessment of reading comprehension assessment, for example, need to evolve over the coming years from information extraction to include information orchestration?

Until now, reading comprehension has been handled by us, human beings, alone. We can see that AI is changing the way we work with language in a fundamental way. In the traditional reading construct, a reader encounters a



text, processes lexical items, builds mental models, makes inferences, and critically evaluates content, all within an individual's cognitive architecture.


In an AI-enhanced reading construct, a reader works with AI to rapidly process, cross-reference, synthesise, and evaluate information from multiple sources simultaneously. The cognitive load shifts from decoding to directing and critically assessing AI-mediated interpretations. This is clearly a different construct. If our testing processes do not change to take this sort of thing into account, we won't be assessing what people actually do in the tests that we use.

For example, when a test-taker can use AI to generate written responses, the relationship between test performance and the underlying construct becomes tenuous. Have you noticed how people's apparent writing ability has changed in the last year or so? I certainly have. In the past I found myself spending a lot of time correcting what people wrote when they were doing minutes, writing reports and so on. There were always issues with grammar, structure, argumentation, vocabulary etc. Then suddenly, when AI was able to transcribe recorded meetings and write minutes, I stopped having to correct minutes. The language proficiency of the people responsible for the minutes and reports hadn't changed. AI had taken over. Two of the most influential language testing researchers of recent years, Bachman and Palmer's (2010), proposed a model of communicative language ability that emphasises the integration of language knowledge with strategic competence. This becomes difficult to assess accurately when AI generates or substantially enhances responses. Writing presents a complex challenge.

If we accept AI as part of communicative competence, writing ability could evolve in a couple of directions:

1. Writing becomes the ability to effectively prompt, guide, and refine AI output to achieve communicative goals. The testing focus shifts from text generation to text orchestration - knowing how to get AI to produce appropriate content and how to shape it for specific audiences and purposes.
2. Alternatively, we develop multiple constructs: "independent writing ability" (purely human) and "AI-assisted communicative competence" (human-AI collaboration), each serving different purposes and contexts and each assessed in different ways.

So, we might need at least two assessment levels:



**Level 1: Core Human Competencies** represent the foundational language abilities that remain necessary regardless of technological context. These include basic linguistic knowledge, fundamental comprehension processes, and core strategic competencies. These competencies align with traditional assessment constructs and remain essential for effective communication in any context and do not change assessment practices.


**Level 2: Collaborative Intelligence** represents emergent competencies that arise from effective human-AI collaboration. This includes the orchestration of AI capabilities to achieve communicative goals that exceed individual human capacity, the development of new discourse conventions for human-AI interaction, and the creation of hybrid communicative competencies that leverage both human and artificial intelligence. This involves the development of new assessment practices reflecting this emerging new construct.

These two levels require multiple construct definitions serving different purposes. Universities evaluating academic writing readiness might focus primarily on Level 1 competencies, ensuring that students possess the foundational abilities necessary for independent scholarly communication. Employers assessing workplace communication skills might emphasise Level 2 competencies, recognising that professional contexts increasingly involve human-AI collaboration. The distinction echoes that made in the 1980s by Cummins (1979) between BICS (Basic Interpersonal Communicative Skills) and CALP (Cognitive Academic Language Proficiency).

Immigration authorities evaluating language proficiency for social integration might require demonstration of core human competencies while acknowledging that settlers will likely use technology-enhanced communication in their daily lives. Different stakeholders can thus receive different but related information from stratified assessment approaches.

So, the challenge relates to the construct definition itself. Traditional definitions of language proficiency have emphasised individual cognitive and linguistic capabilities. However, in a world where AI-assisted communication becomes normalised, language proficiency construct itself is likely to need reconsideration.

For large-scale international assessment providers, these challenges suggest a need to reexamine what they are measuring. Bachman and Palmer (2010) outlined how test usefulness depends on several qualities including authenticity,



reliability, and construct validity - all dimensions affected by AI use. Their framework reminds us that assessment must reflect the target language use domain, which itself is evolving as AI becomes more integrated into communication practices.


Does the rapid development in AI tools mean that we are on the brink of a completely new way of testing that is not about responding correctly but rather about tracking the process needed to respond correctly and deploying technology to do so? AI's potential to identify and track cognitive processes represents perhaps the most transformative possibility for language assessment. Rather than simply measuring outputs, we could assess the underlying cognitive architecture that generates performance. Understanding processes required to generate performance provides much richer information about likely performance across different contexts than any single performance sample.

## **Fairness**

The global scope of major international language tests makes fairness considerations particularly complex when it comes to AI even though many of these considerations already exist. Large assessment providers operate across diverse economic, technological, and educational contexts from major urban centres with advanced infrastructure to remote regions with limited connectivity. As Kunnan (2018) who has written extensively in his research on the concept of fairness in language testing, has long argued, fairness in language assessment requires providing equal opportunities for all test-takers to demonstrate their abilities.

The integration of AI into test preparation (and potentially test-taking) raises significant fairness concerns across all skill areas. Technological access and literacy are not distributed equally across socioeconomic groups. Technology can either mitigate or exacerbate existing inequalities depending on implementation.

For objectively marked reading and listening tests, AI tools can potentially provide advantages through enhanced test preparation, automated analysis of practice tests, and even real-time assistance in identifying patterns or extracting information. Khalifa and Weir (2009) emphasised how familiarity with test formats and strategies affected reading test performance; AI tools that provide sophisticated practice opportunities could create significant advantages for those with access.




More than twenty years ago, Shohamy (2001) argued in her influential work on the power of tests that language assessments already function as gatekeeping mechanisms with profound social consequences. AI technologies have the potential to make these mechanisms even more opaque and potentially unjust if they are not implemented with careful attention to fairness considerations. This is a particularly foggy scene.

## Security

For objectively marked reading and listening tests, security challenges focus on preventing information retrieval and answer sharing. AI tools capable of rapidly retrieving information, recognising patterns in test questions, or even memorising answers from previous administrations pose significant challenges for high-stakes assessments. The implications for objectively marked tests are particularly concerning because these tests often rely on item banks that are reused across administrations. Item harvesting - the systematic collection of test items for distribution or sale – is much easier using AI tools. Such technologies can potentially accelerate the harvesting process by automatically recording, categorising, and storing test content. For international testing organisations, this represents both a security threat and a financial challenge. If item harvesting is successful without AI, just think what can be done with it!

LanguageCert has developed a number of ways to use AI to address security challenges in both high stakes test centres and OLP. These include enhanced identity verification, response pattern analysis, keystroke monitoring, similarity detection in writing tasks and deep fake detection in speaking tests to mention but a few. The challenges nonetheless continue to increase. It would make sense for language test providers to share their information in this area.

On the other hand, AI has already made real time item generation possible although it needs to be improved but we can envisage a time, in the not-too-distant future, (a year or two at most would not appear unrealistic) where AI creates not only test items, but creates unique test items for each administration making memorisation unfeasible. Similarly, it can create infinite variations of core item types measuring at appropriate difficulty levels. Getting the right difficulty level as well as the right content is very important. Auto generation of test items is already a reality and although imperfect now, with continuously refined prompting it will be in common use relatively soon. Our experience at LanguageCert is relevant in this area and may apply to others in international



language assessment. Where 12 months ago auto generation of test items was fully successful only 4% of the time, this year it has risen to around 50%. AI technology used in the production of good items both from the perspective of content and difficulty will be a reality quite soon. It is interesting to consider the effect such progress will have on language assessment.

## **Adaptation Strategies**

Several adaptation strategies are emerging. First, assessment providers are having to reconceptualise what they measure.


The cognitive processes and knowledge structures needed for success in target language environments, increasingly include technology-mediated communication. With objectively marked reading and listening tests for example, this reconceptualisation might focus on higher-order cognitive processes that remain distinctively human. Khalifa and Weir (2009) distinguished between careful reading processes (extracting complete meanings) and expeditious reading processes (quick, efficient information retrieval) - with AI potentially better at the latter.

Test design will need to evolve to incorporate technology explicitly. This requires viewing technology as a component of contemporary language use and hence communicative competence. Just as language tests in the past evolved to incorporate authentic materials and communicative tasks, tests of the future might need to evolve to incorporate AI-mediated communication as part of the assessed construct. For example, we might set writing tasks in an AI simulated environment where the task is not to do the writing per se but to adapt the AI generated output to satisfy relevant communication needs.

Automated test question generation will become widely used, either by allowing for the creation of enormous quantities of test materials or by dynamically creating test items in both computer adaptive and linear contexts we will be able to maintain security (through larger item pools and reduced item exposure) while providing more precise measurement and possibly more tailored assessment.

## **Future Research**

As we navigate these challenges, several research directions emerge and perhaps some of you will investigate these with us.



First, we need systematic investigation of how AI affects test performance across different skill areas and proficiency levels. While theoretical analyses provide important frameworks, empirical research is essential for understanding the real-world impacts of AI on assessment outcomes. Controlled studies comparing AI-assisted and independent performance across different test formats and populations would provide valuable evidence to inform adaptation strategies.

Second, we need validation research for innovative assessment approaches designed to maintain validity in AI-enhanced environments. New item types, integrated task formats, and technology-mediated assessment approaches require validation studies to ensure they effectively measure the intended constructs and provide fair opportunities for diverse test-taker populations.

Third, cross-cultural research on attitudes toward AI in assessment contexts is essential for understanding how different cultural perspectives might influence the acceptance and implementation of various adaptation strategies.

International testing organisations serve diverse global populations; understanding cultural variations in technology acceptance and ethical perspectives is crucial for developing appropriate policies. The only thing is, given the pace at which AI technology is moving forward, will we have enough time for validation research, or will AI do the validation research for us?

## **Conclusion**

In this paper, I have outlined how AI might impact large-scale international language assessment across all skill domains. While the challenges are substantial, thoughtful adaptation rather than resistance offers the most promising path forward. The future of international language assessment depends on our ability to evolve while maintaining the core principles that have always guided our field.

This suggests a couple of ways of looking at the future:

## **Parallel Constructs**

We maintain both "traditional language proficiency" (purely human) and "AI-enhanced communicative competence" as separate, valid constructs serving different purposes. Universities might require traditional proficiency for academic writing, while employers might value AI-enhanced competence for workplace communication.

## **Hierarchical Integration**

We develop a model where traditional skills form the foundation that enables effective AI collaboration. You need strong reading comprehension skills to effectively evaluate AI summaries; you need writing ability to craft effective prompts and refine AI output. AI competence becomes an advanced layer built on fundamental skills. The integration of the two levels is then addressed in relation to user needs.

## **Paradigm Replacement**

We fully embrace AI as part of language competence and redesign assessments around human-AI collaborative communication. Traditional skills like independent reading comprehension become as obsolete for assessment as handwriting has largely become for writing assessment.

As we stand at this crossroads, it is clear that international language assessment cannot remain static. Whether we choose to preserve traditional constructs, layer AI competence onto foundational skills, or redesign our frameworks entirely, the decisions we make now will shape the integrity and relevance of language testing for decades to come. The real challenge is three-fold: technical, ethical and educational - ensuring that assessment continues to empower individuals rather than exclude them, that it reflects real-world communicative practice without sacrificing fairness, and that it keeps pace with the speed of technological change without abandoning the rigorous validation that underpins its credibility. If we succeed, AI will not diminish the value of language assessment but instead expand its scope, helping us to capture more fully the complex, evolving ways in which humans use language to learn, to work, and to live together.

## References

- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.  
<https://doi.org/10.7916/D8CV4HB8>
- Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, interpersonal language skill and bilingual education. *Working Papers on Bilingualism*, (18), 197-205.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.
- Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge.  
<https://doi.org/10.4324/9780203803554>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.
- Microsoft. (2024). *2024 Work Trend Index: AI at work is here. Now comes the hard part*. Microsoft Corporation. <https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part>
- PwC. (2024). *Global Workforce Hopes & Fears Survey 2024*. PwC.  
<https://www.pwc.com/gx/en/issues/workforce/hopes-and-fears.html>
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Routledge.  
<https://doi.org/10.4324/9781315837970>.



# MILLIONS OF EXAMS DELIVERED WORLDWIDE

LANGUAGECERT is an Awarding Organisation recognised by Ofqual. It spearheads innovations in language assessment and certification, providing high-quality services to the global learners' community. It is a UK-based member of PeopleCert Group, a global leader in the certification industry, that delivers millions of exams in over 200 countries.

Learn more about LANGUAGECERT exams at:  
**[www.languagecert.org](http://www.languagecert.org)**

LANGUAGECERT is a business name of PeopleCert Qualifications Ltd, UK company number 09620926