Discussion of AI Landscape in Large-Scale Assessment

March 6, 2025

Goal

The purpose of this session is to provide members with an opportunity to discuss the status of AI use in international and state assessment systems, and to consider how these uses may result in faster, cheaper, and/or better assessment systems. A summary of AI usage in these settings is described below. Board members will be invited to comment on these uses, and to share any additional uses of AI they are aware of that have the potential to lead to faster, cheaper and better large-scale assessment.

Background

AI and International Educational Assessments

The Governing Board has received past presentations from staff with the Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA). Board staff have confirmed that the information received from these sessions is up to date.

TIMMS and PIRLS: The Board learned of ongoing activities for TIMSS and PIRLS in August 2023. These programs have begun to use and explore automated scoring of open-ended responses and automated item generation. The AI supports human scoring of open-ended graphical and written responses and human item writing by drafting items using generative language models and image generators.

For those interested in findings from automated scoring studies for TIMSS, see a recent publication in Science Direct by Jung, Tyak, and von Davier (2025), Towards the implementation of automated scoring in international large-scale assessments: Scalability and quality control.

PISA: As shared with the Board in November 2024, the 2025 PISA assessment will emphasize AI literacy, focusing on cognitive, emotional, and social skills to assess, analyze, evaluate, create, reflect, and engage with online media. Students will use tools like tailored browsers, email, chat applications, and social media to simulate real-world tasks, such as compiling information or evaluating chatbot responses.

In 2025, analytics using AI will be used to investigate learning strategies by analyzing students' processes on PISA. Students will analyze data, conduct experiments, and develop computational artifacts, while the assessment evaluates their persistence, motivation, task engagement, self-reflection, and progress monitoring. Tasks will be open-ended, challenging, and cater to a range of abilities, focusing on the process rather than just correct answers. Post assessment analysis will refine the assessment process based on student reflections on their motivation and feelings about the tasks.

A pilot study of these uses of AI in PISA was conducted in March 2022, and a larger pilot held in five countries in 2023. Additional studies are planned for 2025 and 2026, and if successful, will go operational after.

AI in State-level Assessments

As reported by TeachAI, 26 State Departments of Education have published guidance for using AI in educational settings. These guidance documents primarily focus on teaching and learning and when assessment is described, the focus is primarily on formative assessment (i.e., classroom assessment designed to provide ongoing information about student learning). These guidance documents do not focus on large-scale summative state assessments (e.g., the state assessments used for federal and state accountability). These guidance documents illustrate states are approaching AI with optimism, but also caution. North Carolina's guidance cautions that generative AI should only be used in formative assessment, noting that "Large Language Models and other generative AI tools are new technology and not completely reliable, therefore should not be used to assign letter or number grades to student work."

Though states may be cautious about using LLMs and generative AI for large scale assessment, they have been exploring the use of AI for scoring of constructed response items for some time. A 2013 report presents that the Smarter Balanced assessment consortium began studying automated scoring of constructed response items for more than a decade ago (note that individual states administering Smarter Balanced assessments conduct scoring independently using their own selected processes). States across the country are in various stages of exploring and implementing AI for this purpose.

Texas used automated scoring operationally for the first time in 2024 for constructed response items on their State of Texas Assessments of Academic Readiness (STAAR®). They made this change to allow more constructed response items without increasing the costs and time required to do hand scoring. In April 2024, the Texas Tribune reported that use of automated scoring reduced the number of human scorers required from 6,000 to fewer than 2,000, and saved the state more than $15 million. Human scorers were involved for responses in which the AI program had low confidence.

Beyond automated scoring, an interesting and novel potential use of AI in large scale assessment comes from the Hawaii State Department of Education which issued an

RFP in 2023 that was awarded in 2024 to explore use of AI to enhance efficiency of the assessment development by generating and using virtual students, teachers, and community members in the test development process. If successful, AI could reduce the burden on students and teachers by limiting or eliminating the need for participation in pilot and field testing activities, and reducing the number of teachers and community members required to ensure fair and reliable outcomes.

Definitions

The *Removing Barriers To American Leadership In Artificial Intelligence* Executive Order issued by the White House on January 23, 2025, defines *artificial intelligence* as follows:

> The term *"artificial intelligence"* means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to-
>
> (A) perceive real and virtual environments;
>
> (B) abstract such perceptions into models through analysis in an automated manner; and
>
> (C) use model inference to formulate options for information or action.

Further, *generative AI* is defined by IBM[1] as:

> The term "*generative AI*" means AI that can create original content—such as text, images, video, audio or software code—in response to a user's prompt or request.

[1] What is Generative AI? | IBM