Editorial:
Reflection on the Role of AI in Measurement and Assessment

Hak Ping Tam
National Taiwan Normal University

Ready or not, we are ushered into the era of artificial intelligence (AI), albeit at the dawn of this era. In view of the current interest and speed of advancement in information technology, it is expected that the development of AI will gain high momentum, and its presence will be more obvious in many walks of life in the near future. Not only is AI applied to such commercial and technological areas as robotics and the Internet of Things (IoT), but we have also seen an increasing interest in applying AI to education, from curriculum development to classroom assessment. For example, text-to-video tools can turn a script into a video for instruction with few editing skills. With these tools, classroom teachers are now a step closer to the goal of catering personalized instruction to individual students.

Interest in AI is also on the rise in educational measurement and assessments. In their charting out of the future of assessments, Kyllonen and Sevak (2024) included a discussion of AI and technology-enabled advances in several areas of measurement, including the development of devices that facilitate the inputs and outputs of testing, automatic scoring of constructed-response items, test assembly and generative AI for automated item generation (AIG). There are, however, several other potentially useful applications not mentioned in Kyllonen and Sevak (2024). For example, it is well known that rigorous pilot testing of newly developed items is very costly for testing agencies. Maeda (2024) suggested an approach to substitute human examinees with artificially intelligent examinees. The new approach via AI, if demonstrated successful by further studies, will be much welcomed by testing agencies and authorities in saving developmental time and cost in recruiting a large sample of human examinees and administering the pilot tests to them. The use of AI to bring down costs and reduce production time is a common goal pursued by researchers working on several practical problems in measurement and testing. Regarding bringing down the production cost, one must, however, put into perspective the realistic setup cost of acquiring the services of AI agents as well as the training cost of getting these agents to work (Strubell et al., 2019).

Despite the current high interest in exploring various potential applications of AI, there remain concerns regarding the research of AI in many disciplines. For example, Friedrich et al. (2022) and Láinez-Moreno et al. (2024) both offered suggestions regarding how studies in AI can learn from the history of statistics. Though presented as a document of the Duolingo English test, Burstein (2023) expounded on validity and reliability, fairness, privacy and security, and accountability and transparency as the four responsible AI standards that are applicable to the field of measurement at large. The purpose of this short editorial is to draw the attention of the measurement and testing community to other AI-related issues and offer some suggestions for the future direction of research.

First and foremost, it is advisable to set a high priority on clarifying the nature of AI to ensure its healthy development in matters related to testing and measurement. Artificial intelligence is introduced as "the study of agents that receive percepts from the environment and perform actions" by Russell and Norvig (2021). Many studies in AI paid more attention to the technical aspect of matching actions to the perceived information to maximize the success rate in attaining the study goals. This is, however, the engineering mentality. What kind of intelligence is there in artificial intelligence? Are any of the existing theories of intelligence applicable to AI? Does its intelligence rest in its algorithms, transformers, large databases, or elsewhere? While AI agents are programmed to be capable of learning and reasoning, they do not have a strong self-identity, which indicates intelligence. Without a clear identity of AI, the body of knowledge embodied in and presented by AI agents may amount only to a vast information warehouse and a fast-growing set of models, tools, and orchestration layers to solve practical problems rather than a reliable body of knowledge and wisdom to guide human decision making and resolve dilemmas.

A second issue that is closely related to the first is that for many AI agents, whether traditional AI or generative AI, the way they operate appears as black boxes to their users. AI tools are powerful in the sense that they can learn. However, what they learned is burned in the networks and not revealed to the users (Castelvecchi, 2016). In statistical analyses, the models being used are clearly defined with parameters, and they are fitted to data so that residuals can be assessed. Accordingly, diagnostics can be executed, and model selection can be done. It is unclear, however, the meaning of models adopted in AI tools. The language models or neural network models behind AI are different kinds of models from those adopted in statistics. Their model complexity contributes to AI's image as being mysterious yet magical. There are arguments that we should embrace these black boxes. For example, one can argue that though our brains function as black boxes, we can still trust our brains without knowing how they work. Likewise, we can trust AI as well (Castelvecchi, 2016). However, there are a number of differences between AI and human brains. AI depends on the training data fed into them, yet the sources of the training data, no matter how large, can be biased and hence limited (Domingos, 2012; Láinez-Moreno et al., 2024). Human beings, in contrast, can ask themselves questions, seek opinions from opposite directions, perform self-evaluation, weigh according to ethics, execute remedial action, and apologize if necessary. Our brains may be a black box at some level, yet we can explain to others the logic we reason with, the ethics we abided by, the cultural and religious perspectives we are raised in, and even the emotional influences that we have experienced, rather than based on the probability of words that appear together as in large language models. These and other abilities contribute to our relatively greater trust in our brains.

Let us next turn to several possible directions of future study. First, it is foreseeable that AI will have many applications in measurement and testing. Since it is a quantitative field, it is understandable that researchers pay more attention to the technical aspect of AI when applying it to this area rather than on the philosophical underpinnings of AI as discussed above. Hence, it is suggested that a higher priority be placed on understanding the nature of AI and its models. From a technical perspective, one way to do this is by reverse engineering to open up the black box. More effort in developing explainable AI (XAI) is advisable to assist users in understanding the logic being relied on by AI in making decisions. With better transparency, AI can reliably augment human intelligence to solve complex problems.

Second, there is currently quite some interest in applying AI to facilitate automatic item generation (AIG). At the school level, there is a viewpoint of providing authentic items for students to apply their knowledge in a more realistic way. Authentic situations are usually complex situations. The corresponding test items will probably involve a combination of text, images, tables, graphs, or audio messages. While neural networks can learn features for a single modality, multimodal attentional neural networks (MNN) will need to be developed in order to handle cross modalities feature learning. They will probably facilitate the automatic generation of test items with different stimuli in their item stems or options. Likewise, MNN can also facilitate automatic scoring of students' responses that involve multiple modalities.

Third, regarding generating data by AI for item assessment, effort should be invested in researching how AI can model item-student interactions, including under those situations with unanticipated sources of unfairness. Without knowing in advance what needs to be modeled, how can sources of bias be determined and evaluated? Even if we know those sources, how can their impacts be learned in a measurement and testing situation where the impacts must be known in the first place in order to model them? Items will look as good as they are modeled to look if the data used to evaluate them is from known models (W. Schafer, personal communication, 2025).

Fourth, in view of the current status of technology, it is advisable to regard AI as augmented intelligence rather than as artificial intelligence. It may be more suitable, at least for now, to let AI help us to work smarter rather than doing work for us, especially when educating students are concerned. Presently, many AI researchers are trying hard to build bigger models by introducing more parameters and training them to score high on various tests and examinations. It may be valuable and worthwhile to turn around and train suitable AI agents to prepare students for various assessments at the item, subtest, or whole test levels (W. Schafer, personal communication, 2025). It can help at least those students who are capable but not good at taking tests.

Finally, we turn our attention to a specific direction in AI ethics. It is acknowledged that there are already a couple of documents that discuss areas of attention pertaining to AI ethics on matters related to measurement and testing (e.g., Burstein, 2023). We would like to point out that there is an example from more than 1,500 years ago that we may want to follow. Toward the end of his life, Augustine of Hippo (427/1999) reviewed and commented on his own works, clarifying and self-correcting himself when necessary. This effort was subsequently compiled into the work *The Retractations*. It is suggested that researchers can develop a self-evaluation system in their AI tools similar to Augustine's approach. This is essential since AI can easily generate many new items or automatically score students' answers from tests. The ability to self-evaluate and remediate can increase our confidence in the reliability of decisions from AI agents.

References

Augustine, of Hippo, Saint. (1999). *The retractations* (M. I. Bogan, Trans.). Catholic University of America Press. (Original work published 427)

Burstein, J. (2023). *The Duolingo English test responsible AI standards* [Updated March 29, 2024]. Duolingo. https://go.duolingo.com/ResponsibleAI

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20–23.

Domingos, P. (2012). A few useful things to know about machine learning. *Communication of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Ickstadt, K., Kestler, H. A., Lederer, J., Leitgöb, H., Pauly, M., Steland, A., Wilhelm, A., & Friede, T. (2022). Is there a role for statistics in artificial intelligence? *Advances in Data Analysis and Classification*, *16*, 823–846.

Kyllonen, P., & Sevak, A. (2024). *Charting the future of assessments*. ETS Research Institute. Retrieved from https://www.ets.org/Rebrand/pdf/FoA_Full_Report.pdf

Láinez-Moreno, D., Lorenzo-Arribas, A., & Puig, P. (2024). What can AI learn from the history of statistics? *Significance*, *21*(5), 32–35. https://doi.org/10.1093/jrssig/qmae077

Maeda, H. (2024). Field-testing multiple-choice questions with ai examinees: English grammar items. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/00131644241281053

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics.