Plenary Roundtable Discussion Report: Challenges and Opportunities in Objective Measurement in the Age of Artificial Intelligence

Che Yee Lye[1]

[1] Teaching & Learning Centre, Singapore University of Social Sciences, 463 Clementi Road, Singapore 599494
cylye@suss.edu.sg

Moderator:
Iris Lee, Ministry of Education, Singapore

Panellist, listed in alphabetical order:
George Engelhard, Jr., The University of Georgia, USA;
Guanzhong Luo, South China University and Jiangxi Normal University of China, China;
Jeffrey Durand, Toyo Gakuen University, Japan;
Mohd Zali Mohd Nor, New Star Agencies Sdn. Bhd., Malaysia;
Quan Zhang, City University of Macau, SAR China; and
Zi Yan, The Education University of Hong Kong, Hong Kong SAR, China.

Abstract. Artificial intelligence (AI) and emerging technologies are reshaping educational assessment and raising new questions about the future of objective measurement. This plenary roundtable brought together perspectives from the fields of educational measurement, adaptive testing, language testing, data science, psychometrics and industry to examine key challenges and opportunities in increasingly AI-enabled environments. The discussion highlighted the potential of AI to support automated scoring, item generation, adaptive assessment, personalised learning pathways, real-time feedback and data-informed decision making, while also drawing attention to concerns related to bias, data provenance, privacy, security, transparency and accountability on automated processes. A recurring theme was that technological innovation does not diminish the importance of measurement principles; rather, greater complexity and automation increase the need for clearly defined constructs, valid interpretations, fairness, model fit and invariance. The roundtable also considered practical challenges in translating measurement theory into usable tools, the complementary roles of human expertise and AI, and the ethical implications of AI in both assessment production and evaluation. The discussion suggested that future objective measurement lies not in replacing established measurement principles or human expertise, but in integrating AI responsibly within measurement systems that remain interpretable, defensible and responsive to educational purposes.

Keywords: Objective measurement, artificial intelligence, Rasch measurement, human-AI collaboration.

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_8

1 Introduction: Objective Measurement at an AI Crossroads

The plenary roundtable, “The Road Ahead: Future Challenges and Opportunities in Objective Measurement”, considered how artificial intelligence (AI), digital technologies, adaptive systems and increasingly complex data environments are changing the conditions under which measurement is designed, implemented and interpreted. The discussion was cross-disciplinary, drawing on perspectives from educational measurement, adaptive testing and learning, language testing, statistics and data science, psychometrics, open-source information and industry. This breadth was important because AI does not enter measurement at only one point. It can influence the construction of tasks, the generation of responses, scoring, adaptive delivery, feedback, reporting and decisions based on assessment evidence.

The discussion was framed around several fundamental questions: whether AI should be understood as inherently beneficial or problematic; what measurement integrity should mean in an AI-augmented environment; whether objective measurement changes conceptually when AI becomes embedded in assessment processes; how fairness and accountability can be maintained; and what new skillsets and forms of collaboration will be needed. These questions moved the discussion beyond a simple debate about whether AI should be adopted. A more important issue was where, why and under what conditions AI contributes to objective measurement.

Across the session, the panellists neither uniformly celebrated nor rejected AI. Instead, a conditional position emerged. AI may increase efficiency, expand assessment formats, support personalisation, generate and score assessment content, process large quantities of information and assist human decision making. At the same time, these possibilities introduce risks involving bias, inaccurate information, opaque decision processes, data provenance, privacy, security, overreliance on automation and inappropriate substitution of human judgement. Recent work in educational measurement similarly argues that the introduction of generative AI does not remove established concerns with validity, fairness, security and scale maintenance; it makes them more difficult to ignore [1, 2].

A central proposition running through the roundtable was therefore that technological advancement does not reduce the need for measurement theory. If anything, the growth of AI-mediated assessment increases the importance of clearly defining the construct, specifying the evidence required, evaluating model fit, examining fairness and determining whether score interpretation remains stable under changing conditions. These themes are synthesised in this report through five areas: next generation assessment, AI-enabled applications, the translation of measurement theory into usable practice; measurement integrity; and human-AI collaboration.

2 Next Generation Assessment: Personalisation, Adaptivity and Invariance

One contribution provided a clear vision of next generation assessment through four interacting features: personalised assessment, technology-based assessment, assessment systems supported by generative AI and digital environments [3]. The significance of this vision extends beyond transferring conventional tests from paper to screen. Digital environments make it possible to vary task selection, collect different forms of evidence, respond to learner performance in real time and connect assessment more closely with instruction.

The presentation organised this future around three enduring questions: What does the learner know? What can the learner do? What should the learner do next? The third question is especially consequential. Conventional assessment often ends with a score, grade, proficiency classification, or certification decision. Adaptive and AI-enabled systems can potentially use measurement evidence to determine what learning experience should follow. Assessment then becomes part of a dynamic cycle in which performance generates evidence, evidence informs interpretation, interpretation shapes the next task or learning activity, and the subsequent response generates new evidence.

The discussion of adaptive learning and testing illustrated this possibility. An adaptive educational environment was described as requiring more than an algorithm for selecting the next question. Adaptive testing already provides a technical foundation for selecting items according to estimated ability while preserving measurement efficiency [4]. The new challenge is to extend adaptivity beyond item selection to feedback, sequencing, support and learning recommendations. Although this is pedagogically attractive, it also increases the consequences of measurement error. If a system incorrectly identifies a learning gap, it may direct a learner towards inappropriate content. If the assessment tasks are poorly aligned with curriculum intentions, faster analysis simply enables the system to act on the wrong information more efficiently. The value of personalised assessment therefore depends on the quality of the measurement chain from construct and curriculum through evidence, interpretation, and action.

This is where invariance becomes especially important. In Rasch measurement, invariance represents the aspiration that comparisons should not depend entirely on the particular sample or item set used [5, 6]. The panellists repeatedly returned to the idea that useful measures should remain sufficiently stable when circumstances change. This principle becomes more important when different learners receive different tasks or follow different pathways. The surface experience may be personalised while the meaning of the resulting measures must remain coherent enough to support defensible comparison and decision making.

The roundtable also highlighted an important distinction between prediction and measurement. AI systems may be highly effective at predicting a likely response,

classifying an essay, or recommending an activity. Yet predictive success alone does not establish that the system is measuring the intended variable. A high-performing model may exploit features that are irrelevant to the construct, and a personalised recommendation may be accurate on average while functioning poorly for particular groups or response types. Measurement theory therefore contributes questions that computational optimisation alone does not answer: What variable is being represented? What evidence supports the inference? How much uncertainty remains? Does the relationship hold across persons, tasks, and contexts? These questions remain central as assessment becomes increasingly adaptive and AI-enabled.

3 AI-enabled Assessment Applications: Item Generation, Scoring and Language Testing

The roundtable considered several uses of AI in assessment, including automated essay scoring, topic modelling, automated item generation, difficulty prediction, assessment security and adaptive delivery. These applications demonstrate the potential for substantial gains in speed and scale, but the panellists also make clear why efficiency cannot be treated as evidence of measurement quality.

One presentation referred to an automated essay-scoring application achieving approximately 80% agreement or prediction accuracy. Rather than treating that figure as sufficient evidence of success, the discussion focused on the remaining 20%: Which responses are scored inaccurately and why? The same logic applies to automated item generation. Large language models (LLMs) can produce large numbers of plausible questions rapidly, but a generated item is not necessarily a calibrated item. It may be grammatically convincing while differing from intended difficulty, representing the construct poorly, cueing the answer, or functioning unexpectedly for particular learners. The roundtable positioned AI as a potentially powerful component of item development rather than as a replacement for empirical calibration and expert judgement.

A related possibility involved combining empirical, judgemental and AI-enabled approaches to item banking. AI could support the development of assessment items by drafting questions, predicting item characteristics, identifying content coverage or generating alternative versions. Measurement evidence and expert review would be used to evaluate how those items actually function. Such an integrated approach is more defensible than treating item development as either entirely human or entirely automated. The central principle is that AI can expand and accelerate the generation of assessment items, but psychometric evidence and professional judgement remain necessary to determine whether those items can support valid measurement interpretations.

Language testing brought many of these issues together particularly clearly. Traditional language test development can involve substantial work in item writing, review, moderation, standard setting, equating, scoring and reporting. AI has the

potential to accelerate several of these processes through automatic item generation, automated review, multilingual reporting, response-pattern analysis, adaptive testing and real-time feedback. However, the language testing contribution drew a firm boundary between generating content and establishing its measurement properties. AI may propose an item; it does not automatically establish its difficulty, discrimination, fairness or suitability for equating. Human oversight was therefore regarded as necessary, especially for assessment design, item analysis, equating, validation and consequential decisions. The position was not to reject AI, but to adopt it conditionally: AI should be used where it contributes to a clear assessment purpose and where its outputs can be evaluated, challenged and governed.

The discussion also anticipated a future in which assessment becomes less episodic and more deeply embedded in learning. Rather than stopping instruction to administer a separate test, digital environments may continuously collect evidence while learners work through authentic activities. AI may then support rapid analysis and feedback. Such integration has considerable promise, but it also blurs boundaries between teaching, assessment and intervention. If feedback changes subsequent performance, the system must be clear about whether it is estimating current unaided capability, supported performance, learning potential or some combination of these. Again, the construct must precede the technology.

A final concern involved the data used to train AI systems. LLMs may reflect the composition and omissions of the information to which they were exposed. If published literature overrepresents positive findings or if particular perspectives are absent, an AI system may reproduce those imbalances in generated assessment content or recommendations. The implication is that validation must extend beyond the visible output to consider the information environment that produced it. AI-mediated assessment therefore requires both psychometric evaluation of outputs and scrutiny of the data and assumptions embedded upstream.

4 From Measurement Theory to Practice: Statistics, Usability and Industry

A distinctive feature of the roundtable was its attention to implementation. Sophisticated measurement theory has limited educational impact if it remains accessible only to specialists. During the open discussion, a practical gap was identified between learning a measurement procedure and sustaining its use. Classroom teachers can learn the mechanics of a basic Rasch analysis relatively quickly, yet many do not continue using the procedures once formal training has ended. The challenge, therefore, is not simply a lack of measurement literacy, but also of usability and workflow. Specialised software, data preparation, technically demanding interpretation and processes that sit outside everyday teaching can make otherwise useful measurement practices difficult to sustain.

AI may provide a bridge if it can reduce unnecessary technical burden without obscuring the measurement logic. The proposed direction was not merely to add a conversational interface to existing software, but to develop tools in which Rasch or Item Response Theory (IRT) principles are embedded within useful educational workflows. Teachers might receive interpretable information about learning gaps, progress, targeting or item functioning without manually executing every analytic step. However, a responsible system should alert users when data are inadequate, items misfit, uncertainty is large or a proposed interpretation is not justified.

This implementation challenge also has implications for professional capability. Future capacity may depend as much on interdisciplinary communication as on advanced technical knowledge. Psychometricians need to understand how AI systems are built and deployed, while developers need to understand that measurement is more than prediction. Teachers, in turn, need enough measurement literacy to question automated outputs rather than treating them as neutral facts.

Another contribution sought to connect AI with established statistical traditions. Statistics was described through the familiar sequence of specifying a model, estimating parameters, and evaluating fit, while AI was presented as involving models, training data, parameter estimation and iterative adjustment. A provocative analogy suggested that IRT could be viewed as an "AI version" of Classical Test Theory (CTT). This comparison was challenged during the audience discussion and subsequently clarified as an analogy about iteration and computation rather than a literal claim that IRT is AI. The exchange was useful because it highlighted both continuity and difference. AI and psychometrics share concerns with models, estimation, fit, data quality and uncertainty, but objective measurement adds substantive requirements concerning construct meaning and invariance that cannot be reduced to computational complexity [5, 7].

The industry contribution broadened the discussion beyond education. Shipping was used to illustrate AI applications in supply-chain management, demand forecasting, route planning, navigation and cargo safety. One example involved automated screening of cargo declarations to flag potentially mis-declared hazardous goods. The case demonstrated how AI can support high-stakes decisions at scale, but it also reinforced the continuing role of human expertise. AI works most readily with explicit knowledge represented in records, documents and codified procedures; people continue to contribute tacit knowledge, contextual judgement, intuition, empathy and responsibility in complex situations.

The industry perspective also brought practical governance challenges into sharper focus, including privacy, security, data quality, skill gaps and regulatory requirements that directly shape how AI systems can be deployed. Sensitive commercial information, for example, may not be suitable for use with open models, requiring organisations to rely on controlled internal data and more restricted system architectures. It was also noted that Rasch measurement was not currently used in the industry applications described during the roundtable. However, measurement principles were suggested as

a possible future avenue for evaluating or optimising such systems. This possibility should therefore be regarded as an exploratory research direction rather than an established industrial practice.

5 Measurement Integrity in AI-Enabled Systems: Data Provenance, Bias and Accountability

The concept of measurement integrity expanded during the roundtable beyond the psychometric properties of scores to include the provenance and governance of information entering AI-enabled systems. The discussion of open-source intelligence illustrated this concern through the idea of provenance: where information originated, what chain of control it passed through and whether it was manipulated. AI can contribute to the production of convincing fake text, images and video, while false information placed online may subsequently be incorporated into later AI outputs or training data. The relevance to measurement is direct. A sophisticated model cannot automatically repair evidence that is fabricated, contaminated or systematically unrepresentative. In AI-rich environments, measurement integrity therefore reaches upstream to the origin and quality of the observations themselves.

Bias became a related focus in the audience discussion, where questions were raised about the strategies and frameworks needed to mitigate bias in AI-enhanced measurement systems. The responses highlighted the importance of defining bias precisely, because the term carries different meanings across AI and statistical contexts. In AI discourse, bias may arise from training data, model design, representation or deployment, and may manifest as systematically different or unfair outcomes for particular groups [8]. In statistical estimation, by contrast, bias has a more specific technical meaning: the systematic difference between the expected value of an estimator and the parameter it is intended to estimate [9]. These distinctions are important because strategies for addressing demographic or algorithmic unfairness are not necessarily the same as those used to address bias in statistical estimation. Without specifying the form of bias under consideration, both its diagnosis and mitigation risk remain conceptually vague.

From an industry perspective, one proposed strategy was to constrain an AI system to a clearly defined domain and train or configure it with relevant, controlled information rather than allowing indiscriminate reliance on open data. From a statistical perspective, the emphasis was on defining the expected relationship, comparing observed and expected behaviour, identifying systematic discrepancies, and then determining whether adjustment is warranted. The latter logic is familiar to psychometrics through model-fit analysis, residuals, differential functioning and related techniques.

Accountability is the final component of measurement integrity. The roundtable did not produce a single allocation rule, but the contributions consistently rejected the idea

that responsibility can simply be transferred to the machine. If AI generates an item, scores an essay, identifies a learning gap or recommends a pathway, the people and organisations that design, approve and use the system remain part of the decision chain. Responsibility should therefore be made explicit rather than assumed. The greater the consequence of a decision, the clearer the process should be for reviewing automated outputs, challenging errors, documenting the basis for decisions and determining who is accountable.

6 Human-AI Collaboration: Ethics, Expertise and Creativity

The closing discussion moved from technical implementation to a deeper examination of what should remain distinctively human in an AI-enabled measurement environment. Across the roundtable, human expertise was defended for different reasons. In language testing, humans were needed for construct definition, validation, equating, and professional judgement. In industry, people contributed tacit knowledge and contextual interpretation. In classroom practice, teachers needed to interpret information in relation to learners and curriculum. The common point was not that humans are always more accurate than machines, but that measurement decisions involve purposes, values, context and responsibility that cannot be delegated solely on the basis of computational performance.

A concluding contribution introduced a more sceptical perspective on the terminology of AI, questioning whether the term “intelligence” risks overstating the capabilities of current generative systems. These systems were charaterised primarily in terms of probabilistic prediction, while their usefulness for tasks such as language correction and translation was acknowledged. Regardless of whether this characterisation is accepted in full, the intervention served an important purpose by shifting attention from what AI is called to what it actually does and what claims about its capabilities can be justified.

From this perspective, an ethical asymmetry in educational assessment was raised. Institutions may restrict students from using generative AI to produce assessed work while simultaneously using AI to generate questions, provide feedback, or mark responses. This does not necessarily imply that automated marking is unethical or that students should have unrestricted access to AI. Rather, it highlights the need for a coherent construct rationale. If an assessment is intended to provide evidence of independent human capability, then the conditions under which responses are produced and evaluated should support that interpretation. Conversely, if AI-assisted performance is part of the competence expected in contemporary practice, the construct may need to include competent human–AI interaction

The key measurement question is therefore whose performance a score represents. When AI contributes substantially to the production of a response, the observed outcome may represent human-AI performance rather than unaided human capability.

These are not necessarily competing constructs, but they are different constructs and should be distinguished explicitly in assessment design.

The final reflection concerned creativity and discovery. It was argued that important intellectual advances do not always begin with a clearly specified problem and proceed through efficient optimisation. Curiosity, sustained engagement and unexpected connections may reveal a problem only after a possible solution or conceptual insight has emerged. Examples from technological discovery and extended work with Rasch models were used to illustrate this point. The broader concern was that organisations adopting AI primarily to reduce human effort may also reduce opportunities for exploration work from which new ideas and discoveries emerge.

7 Future Directions and Conclusions

The plenary did not prescribe a single model for the future of objective measurement, but several priorities emerged from the discussion. First, technology should follow the measurement purpose rather than determine it. Before an AI tool is introduced, the construct, required evidence, intended interpretation and decision context should be clear. This principle protects assessment from becoming technology-led rather than evidence-led.

Second, AI-mediated assessment should be evaluated using criteria that extend beyond aggregate accuracy. Invariance, model fit, interpretability, fairness, targeting, uncertainty and the behaviour of anomalous cases remain important. In some cases, the failures of an AI system may be more informative for validation than its average success. Understanding which essays are mis-scored, which generated items misfit or which learners receive inappropriate recommendations can reveal construct and fairness problems that headline performance statistics conceal.

Third, data provenance should become an explicit part of measurement integrity. Researchers and practitioners need to know where AI inputs originated, how they were selected, whether they were manipulated and which perspectives may be absent.

Fourth, AI-enabled measurement tools should be designed around the workflows of actual users. For classroom practice, this means embedding psychometric principles in interpretable tools rather than expecting every teacher to become a specialist software user.

Fifth, accountability must remain visible. Systems should specify where automated recommendations end and professional judgement begins, particularly when decisions have meaningful consequences.

Sixth, fairness and bias should be operationalised rather than left as broad aspirations. Psychometric traditions of model fit, differential functioning, residual

analysis and comparison between observed and expected behaviour provide useful starting points for examining systematic discrepancies.

Finally, the development of AI-enabled measurement should preserve human capacities that technological systems are intended to support rather than displace professional judgement, curiosity, responsibility and creativity.

The roundtable portrayed a field undergoing technological changes while continuing to confront questions that are fundamentally about measurement. AI can generate assessment content, support scoring, analyse responses, support adaptive delivery, provide rapid feedback and contribute to decision making. Yet these capabilities do not remove the need to establish what is being measured, whether the observations provide appropriate evidence of the intended construct, whether comparisons are fair and whether interpretations remain sufficiently stable under changing conditions.

In this context, the continuing relevance of invariance is especially important. Objective measurement is not simply about producing numbers efficiently. It is about developing measures whose interpretations can extend beyond the immediate sample, item set or technological environment. AI therefore presents both an opportunity and a test for this aspiration. It can make measurement more scalable, adaptive and responsive, while measurement theory provides the principles needed to judge whether AI-enabled outputs can be interpreted as trustworthy evidence.

The road ahead is therefore not one in which AI replaces established measurement principles, but one in which those principles become even more important. The enduring questions of objective measurement – what the variable is, what observations represent it, what assumptions connect observations to measures, how invariant the interpretation is, where bias enters, and what decisions are justified – are not made obsolete by AI. They become more urgent as assessment systems become increasingly automated, adaptive and technologically complex.

Acknowledgments. The author gratefully acknowledges Professor David Andrich for his contribution as a special guest panellist of the Plenary Roundtable Discussion at the Pacific Rim Objective Measurement Symposium (PROMS) in July 2025, at Singapore University of Social Sciences (SUSS). Appreciation is also extended to the panellists for reviewing the report and confirming the accuracy of its representation of the discussion.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this paper.

References

1. Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. Transforming assessment: The impacts and implications of large language models and

generative AI. Educational Measurement: Issues and Practice, 43(2), 16-29 (2024). https://doi.org/10.1111/emip.12602
2. Ho, A. D. Artificial intelligence and educational measurement: Opportunities and threats. Journal of Educational and Behavioral Statistics, 49(5), 715-722 (2024). https://doi.org/10.3102/10769986241248771
3. Lee, E., & Engelhard, G., Jr. Next generation assessments and Rasch measurement theory. Rasch Measurement Transactions, 37(1), 1956-1959 (2024).
4. van der Linden, W. J., & Glas, C. A. W. (Eds.). Elements of adaptive testing. Springer. (2010). https://doi.org/10.1007/978-0-387-85461-8
5. Andrich, D., & Marais, I. A course in Rasch measurement theory: Measuring in the educational, social and health sciences. Springer. (2019). https://doi.org/10.1007/978-981-13-7496-8
6. Engelhard, G., Jr., & Wang, J. Rasch models for solving measurement problems: Invariant measurement in the social sciences. SAGE. (2021). https://doi.org/10.4135/9781071878675
7. Wilson, M. Constructing measures: An item response modeling approach. Lawrence Erlbaum Associates. (2005).
8. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115, 1–35 (2021). https://doi.org/10.1145/3457607
9. Kosmidis, I. Bias in parametric estimation: Reduction and useful side-effects. WIREs Computational Statistics, 6(3), 185–196 (2014). https://doi.org/10.1002/wics.1296

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

