Viewpoint | Published 14 August 2024 | doi:https://doi.org/10.57187/s.3862
Cite this as: Swiss Med Wkly. 2024;154:3862

Shake-up in the world of assessment: Impressions from the Ottawa Conference on Assessment from Down Under

Raphaël Bonvin[a], Bernard Cerutti[b]

[a] Medical Education Unit, University of Fribourg, Fribourg, Switzerland
[b] Faculty of medicine, University of Geneva, Geneva, Switzerland

Assessment plays a key role in the legitimisation of educational institutions by the community, the professions and policymakers, but it has usually not received the attention it deserves. Ian Hart and Ronald Harden identified a need for a conference on the topic of assessment of clinical competence to facilitate the sharing of views and experiences. They launched the first biennial international Ottawa Conference on Assessment in Medicine and Healthcare [1] in 1985 in Ottawa (hence the name) to bring together the world's specialists, experts and opinion leaders to share advances in the field of assessment in medicine. The 24th edition was held in Melbourne in February this year. What follows is a summary of the key themes and issues that the authors have drawn from the presentations, workshops and discussions.

Artificial Intelligence is everywhere

The major breakthrough of generative Artificial Intelligence (AI) resulted from a long process that began with text analytics and rule-based systems initiated in the fifties and the sixties. However, with its newfound availability and accessibility, generative Artificial Intelligence has abruptly shifted trust from our academic research institutions to the whole educational and public sphere: this requires a rethinking of what should be taught and a rapid transformation of our assessment methods, whether we like it or not.

The fact that general-purpose Artificial Intelligence systems, not specifically trained in medicine, successfully pass medical licensing exams [2] tells us more about how we assess than about Artificial Intelligence itself. The values that underpin today's approaches to assessment are finite knowledge (a static chapter or subject at one point in time), highly controlled learners (we test exactly what has been taught), individual performance, a suspicion-based relationship (teachers anticipate that students will cheat and students look on assessment as a tool to make them fail), linear and predictable outcomes, reductionism (we test what is easy to test and condense it all to information-poor grades). These values are far removed from the professional reality, and make our assessment culture increasingly at odds with the future clinicians we want. Assessment culture should reflect the distributed cognitive system in which students are constantly interacting with peers, teachers and tutors (by the way the Health Professions Artificial Intelligence Tutor already exists) and the whole community, invariably surrounded by generative AI, social media platforms, instant messaging, online encyclopaedias and databases, and so on. Within this general framework, a valid assessment must be meaningful, active and collaborative.

Another aspect of Artificial Intelligence that was discussed was its ability to correct text-based exams, making short-answer questions and essays realistic for large cohorts. Also, the potential of Artificial Intelligence to evaluate all available data generated by learners during their training could support a review and decision process by a competence committee.

Exam formats

There can be no conference on assessment without a reflection on the two main formats: Multiple-Choice Questions (MCQ) and Objective Structured Clinical Examination (OSCE). They are here to stay in one form or another, but they are losing their exclusive role.

Multiple-choice questions (MCQ)

A closed-book exam where students have to rely only on their "biological" memory, isolated from external sources of information and interactions is completely at odds with what is expected of a clinician who looks up and discusses with peers and experts when uncertain. There is no point in investing a huge amount of resources in an exam that a bot will likely pass with a high score! The focus should shift to how to work with, understand and be sceptical and creative about available data and information.

Objective Structured Clinical Examination (OSCE)

Introduced in the mid-seventies, Objective Structured Clinical Examinations aimed to reduce the number of variables affecting performance assessment by increasing standardisation [3]. They are reliable and demonstrate educational impact, but they are also extremely costly. Some see them as an "assessment factory": highly efficient but narrow in scope. Some institutions cancelled their OSCE

Dr Bernard Cerutti, MPH
University of Geneva
Faculty of medicine
UDREM
Rue Michel-Servet 1
CH-1206 Geneva
bernard.cerutti[at]unige.ch

during the COVID pandemic and decided not to reintroduce them afterwards. Their rationale: at the end of the day, OSCEs only prepare students to do well in their final OSCE exams. They tend to induce bad habits of scattergun, robotic, formulaic racing and score chasing. Very different from the professional and compassionate patient care we wish for! New formats have been proposed that differ noticeably from the classic OSCE, such as an authentic and reliable adaptation of the Objective Structured Long Examination Record (OSLER) [4] longer than an Objective Structured Clinical Examination, more authentic (sometimes real patients rather than simulated), more time spent on communication skills and approaching the patient as a whole, and a subgroup of examiners more skilled in assessment and feedback; or the Assessment for Progression Exam (APEX) [5] with a flexible timetable, loose timing (no bell), a feedback phase and trained examiners. But whatever the format, the current technological leap is such that this type of assessment should focus much more on dimensions such as communication (often underweighted in the score calculation), case management and interprofessionalism, which is rarely the case.

Entrustable Professional Activities

Entrustable Professional Activities (EPAs) were the subject of much debate, particularly in postgraduate education, where they are now well established. Evidence of their relevance can be seen, for example, in identifying struggling residents early in their training and giving them the support they need to improve and develop without wasting unnecessary years of training. The difficulties of getting to grips with the concept of ‘entrustment’ were discussed and the different types and approaches of competency committees/entrustment committees were explored. However, the EPA concept seems not always fully understood (e.g. the difference between competency and skills, the necessity of multiple observations from multiple observers in multiple situations, the differences between ad hoc evaluation and entrustment decision), leading to confusion about the soundness of the assessment approach and the reliability of the judgements.

Programmatic assessment

In addition to Artificial Intelligence, Programmatic Assessment (PA) [6, 7] was the other main thread of the conference. From a few undergraduate and postgraduate programmes that introduced Programmatic Assessment a few years ago (including the Master of Medicine at the University of Fribourg [8]), the concept is now spreading to many programmes. It is seen as the answer to many problems, including those raised by Artificial Intelligence. But behind the buzzword, principles of PA are not always respected, and the incompatible *summative game* is still very much dominant. More than just using some assessment formats and labelling assessments as formative, PA is a different ecosystem of values and a different understanding of the roles of teachers and learners. High-stake decisions must be based on triangulation of a rich set of information from multiple sources (e.g. longitudinal data, meaningful feedback on targeted learning activities), which implies a combination across formats, and a narrative rather than a purely numerical process.

Future perspectives for licensing exams

Assessment in medical education is not static and has evolved significantly over the years and decades [9]. The expansion of Artificial Intelligence and the promise of a much more comprehensive – covering all domains of competence – and valid assessment offered by a Programmatic Assessment approach challenges the single-shot format of licensing and certifying examinations (typically administered with Multiple-Choice Questions and an Objective Structured Clinical Examination) in terms of validity, efficiency and educational impact. The idea is that, at least, a short-term change in content must be considered (see discussion above on MCQ and OSCE) and, in the longer term, a combination – some even suggest a replacement – with more longitudinal and comprehensive approaches should be pursued.

Conclusion

Assessment is the Curriculum – this statement fully expresses the importance of assessment in education. It underscores the need to think and build meaningful approaches that support the expected learning to best prepare our students and residents for their professional activities. The Ottawa Conference provided a stimulating, sometimes confronting, moment in looking at what the future of assessment might look like. In any case, old assumptions and habits will certainly be shaken by the changes driven by Artificial Intelligence and many other factors.

Potential competing interests

Both authors have completed and submitted the International Committee of Medical Journal Editors form for disclosure of potential conflicts of interest. No potential conflict of interest related to the content of this manuscript was disclosed.

References

1. https://www.ottawaconference.org
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb;2(2):e0000198. http://dx.doi.org/10.1371/journal.pdig.0000198. PubMed. 2767-3170
3. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach. 2013 Sep;35(9):e1437–46. http://dx.doi.org/10.3109/0142159X.2013.818634. PubMed. 1466-187X
4. Gleeson F. AMEE Medical education guide no 9: Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). Med Teach. 1997;19(1):7–14. http://dx.doi.org/10.3109/01421599709019339. 0142-159X
5. Department of General Practice Annual Report. 2023. School of Public Health and Preventative Medicine, Monash university, Melbourne Austria (2023).
6. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85. http://dx.doi.org/10.3109/0142159X.2011.565828. PubMed. 1466-187X
7. Heeneman S, de Jong LH, Dawson LJ, Wilkinson TJ, Ryan A, Tait GR, et al. Ottawa 2020 consensus statement for programmatic assessment - 1. Agreement on the principles. Med Teach. 2021 Oct;43(10):1139–48. http://dx.doi.org/10.1080/0142159X.2021.1957088. PubMed. 1466-187X
8. Bonvin R, Bayha E, Gremaud A, Blanc PA, Morand S, Charrière I, et al. Taking the Big Leap: A Case Study on Implementing Programmatic Assessment in an Undergraduate Medical Program. Educ Sci (Basel). 2022;12(7):425. http://dx.doi.org/10.3390/educsci12070425. 2227-7102
9. Schuwirth LW, van der Vleuten CP. A history of assessment in medical education. Adv Health Sci Educ Theory Pract. 2020 Dec;25(5):1045–56. http://dx.doi.org/10.1007/s10459-020-10003-0. PubMed. 1573-1677