Australian Medical Council Limited

Evidence from medical education literature - strengths and weaknesses in assessment

The Covid 19 pandemic has had a significant impact on education and assessment in Australia and around the world. Although the disruptions in Australia and New Zealand may have been less than many other places in the world, most organisations in medical education have had to adapt their processes. It is fair to say that these many examples of Plan B solutions have met various degrees of success. With the arrival of vaccinations and good hopes for an end to the disruptions in the near future the question what the 'new normal' will look like and how to prepare for it are both relevant and timely. Our viewpoint is that the 'new normal' will not likely be the same as the 'old normal' and, more importantly, that it should not be the same as the old normal. Amongst other things, Covid 19 has shown that the old normal was likely to be too vulnerable for disruption and not in keeping with the advances in the relevant literature.

Therefore, with Covid 19 as an unexpected 'catalyst' for improvement and change of current assessment processes, it may be wise to consider some of the robust evidence in the medical education literature about strengths and weaknesses around assessment. The most important of these are discussed in this document. Every subsection makes reference to literature. Each reference is only one example of that literature, and each subsection could be supported by many references.

- the issue of adequate sampling

Every assessment is in fact a small sample out of the whole domain of relevant questions, stations, assignments that could have been used. Even a 200 item multiple-choice examination is only an 'n' of 200 out of the domain of at least tens of thousands of relevant possible questions. Like in research, the smaller the study sample, the lower the generalisability of the results to the population at large, and the less the likelihood of reaching any statistical significance. Sampling does not only relate to the number of items in an assessment but also to the number of examiners, stations and even the number of occasions at which the exam took place. An exam that takes place for one day only is likely to be a more limited sample than assessment on a more longitudinal basis. As in clinical medicine, poor use of a diagnostic procedure or inadequate sampling is not only likely to produce false negatives – candidates failing who are actually sufficiently competent – but also to engender false positives - candidates passing who are actually not sufficiently competent. So, any exam that is based on a limited number of cases, includes judgements from a limited number of examiners or involves observations from limited sources on limited occasions, is likely to produce a significant number of false positive and false negative results[1]

- the issue of domain specificity

[1] - Swanson DB. A measurement framework for performance-based tests. In: Hart I, Harden R, eds. Further developments in Assessing Clinical Competence. Montreal: Can-Heal publications 1987:13 - 45.
- Swanson DB, Norcini JJ. Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine* 1989;1(3):158-66.
- Norcini JJ, Swanson DB. Factors influencing testing time requirements for measurements using written simulations. *Teaching and Learning in Medicine* 1989;1(2): 85-91.

Unfortunately, all components of competence suffer from domain (aka content) specificity. This means that performance on one case, station or assignment is a poor predictor of how the same candidate would perform on any other relevant case, station or assignment. This is a counterintuitive concept. We often think that if we have observed a candidate in one situation, we can reliably draw inferences from this and make generalised judgements as to whether the candidate is a competent doctor or not. Unfortunately, this is not the case and is a very robust finding in the literature. The explanation for the phenomenon of domain specificity is quite complex and centres on the capacity of seemingly different cases to connect to the same underlying principle or competence[2]. This has ramifications for generalised judgements about a candidate based on one single observation or case. A candidate who performs poorly on one case and fails an assessment, might have done perfectly on all other given cases, but also a candidate who performs well on a certain given case might have performed very poorly on all other given cases.

- The difference between assessment format and assessment content

Although it is customary in assessment practice to be primarily focused on the format of an assessment, it is actually the content that determines the validity. Counterintuitively, when the same content is being asked of a candidate, the format is relatively unimportant. This has even been demonstrated when comparing an actual, practical OSCE with a written test on physical examination skills[3]. This is probably the most counterintuitive finding and such comparative studies are relatively rare in the literature, but there are myriads of publication comparing different item formats – typically open-ended with multiple-choice– in the medical education literature. In a nutshell, they almost unanimously show that competence does not generalise well across contents but extremely well across formats. So, two multiple-choice items asking different things do not correlate well, and the same holds for two open-ended questions or essays, but a multiple-choice question and an open-ended question asking for the same (applied) knowledge aspect correlate very highly. Therefore, careful item or clinical station writing, thorough review, and post-test psychometric analysis with moderation, contribute more to the validity of an assessment than specific scoring rules, complicated formats and weighting or the way in which numerical scores of different assessments are combined.

– the issue of validity

A central problem in all assessment is the fact that we are trying to assess something that we cannot observe directly. Where, for example, a patient's weight can be both measured but also gauged by observation, every aspect of competence has to be inferred from what is observable. This is a bit like taking a blood pressure. Blood pressure cannot be observed directly, and it has to be inferred from reading a sphygmomanometer whilst gradually lowering the pressure in the cuff auscultating the brachial artery. So, in order to assure that the blood pressure measurement is valid we have to be certain that the measurement is based on a correct procedure, in other words that the observations made by the clinician (from the sphygmomanometer) are correctly translated into numbers. It is also important that sufficient blood pressure measurements are taken to ensure that the findings are reproducible and that the findings correspond with other measures around cardiovascular health (such as pulse, auscultation, jugular venous pressure, et cetera)[4]. Validity in assessment follows a similar

---

[2] - Eva KW, Neville AJ, Norman GR. Exploring the etiology of content specificity: Factors influencing analogic transfer and problem solving. *Academic Medicine* 1998;73(10):s1-5.

[3] - Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. *Medical Education* 1988;22:97-107.

[4] - Llabre MM, Ironson GH, Spitzer SB, Gellman MD, Weidler DJ, Schneiderman N. How Many Blood Pressure Measurements are Enough? An Application of Generalizability Theory to the Study of Blood Pressure Reliability. *Psychophysiology* 1988;25(1):97-106.

pattern; procedures have to be in place to ensure that the observation of performances correctly translate into scores, that the scores are based on a sufficiently large sample to ensure that they are reproducible/generalisable and that the findings correspond with other measures of assessment so that a complete image of a candidate's competence can be validly made[5].

- Reliability

In its classical sense reliability purely indicates the reproducibility of outcomes of an assessment. This means, in its strictest interpretation, that if a candidate obtains a certain score – let's say 58% – he or she should obtain the same score if he or she were tested again with a similar test of similar difficulty. The slightly less strict interpretation is the expectation that the candidate's position in the rank order from best performing to most poorly performing would be the same, i.e. if they were the fourth best performing candidate on the assessment they would be expected to also be the fourth best performing candidate on a similar assessment. This second interpretation is most often used, for example in the rather famous Cronbach's alpha[6].

This straightforward approach to reliability as reproducibility has long been the only one. However, when assessment started to include human judgement more prominently, and with the increased awareness that competence is not something that can only be expressed in scores but also in narratives, other approaches to reliability have since gained importance. One such approach is based on the concept of saturation of information[7]. Although this concept is derived from qualitative research it is also something that is well-known to almost any practising clinician. When conducting a diagnostic workup, there is always a moment at which the clinician decides that no further diagnostic information is needed, because the diagnosis or the preferred management can be determined with sufficient certainty. This too is a saturation of information principle and can be applied in the same way to assessment.

- The role of feedback

There is overwhelming support in the literature that providing constructive and meaningful feedback leads to more rapid development of expertise and, eventually, to higher levels of expertise.[8] Unfortunately, many educational contexts in medicine do not have a culture of providing constructive and meaningful feedback and of 'closing the loop'[9]. It is clear that this can be seen as a missed opportunity because where there are systems of identifying registrars who are struggling and giving them access to feedback and remediation opportunities they are considerably more likely to perform

---

5 - Kane MT. Validation. In: Brennan RL, ed. Educational Measurement. Westport: ACE/Praeger 2006:17 - 64.

6 - Clauser BE, Margolis MJ, Swanson DB. Issues of validity and reliability for assessments in medical education. In: Holmboe ES, Hawkins RE, eds. Practical Guide to the Evaluation of Clinical Competence. 1st ed. Philadelphia: Mosby/Elsevier 2008:10 –23.

7 - Driessen E, Van der Vleuten CPM, Schuwirth LWT, Van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Medical Education* 2005;39(2):214-20.

8 - Ericsson KA, Charness N. Expert performance. *American Psychologist* 1994;49(8):725-47.
- Ericsson KA. An expert-performance perspective of research on medical expertise: the study of clinical performance. *Medical Education* 2007;41:1124-30. doi: 10.1111/j.1365-2923.2007.02946.x

9 - Watling C, Driessen E, Van der Vleuten CPM, Vanstone M, Lingard L. Beyond individualism: professional culture and its influence on feedback. *Medical Education* 2013;47(6):585-94.

well. For example, on the fellowship examinations[10] . The incorporation of feedback cycles, focusing on strengths but also weaknesses in combination with opportunities to practice and improve the weaknesses or to retain the strengths with repeated observation, is often called 'deliberate practice'[7].

- The role of the supervisor or assessor

Whereas in written or computerised assessment, validity can be built into the assessment through careful test production, this is not the case with workplace based assessment. In workplace based assessment, the quality of the assessor – their ability to translate what they observe into a meaningful result or score – is essential for validity. Untrained assessors will not be able to produce high-quality assessment results. Structured rubrics may mitigate this negative effect of lack of training of assessors[11], but only to a small extent[12]. An important implication of this is that a comprehensive 'picture' of a registrar's or candidate's competence can only be obtained when multiple stakeholders are involved. Each stakeholder has expertise to see certain aspects but may be blind to others. For instance, a scrub nurse may not be a good person to ask about a surgeon's interaction with patients, but may know a great deal about their sensitivities and respect for tissue, and they have far more experience with a range of surgeons. This is the reason why instruments such a multisource feedback are a valuable addition to the range of instruments in an assessment program.

Another development that has demonstrated its usefulness in supporting the assessor in making valid decisions is the use of so-called entrustable professional activities (EPAs)[13]. The biggest advantage of EPAs is that they employ a language which is more intuitive to most clinical supervisors. This is certainly not trivial. One could argue that by asking supervisors to use judgements they have more experience with, instead of using more 'educational' language, they are actually put in a more 'expert' position. Good EPAs lead to demonstrably positive effects on the quality/validity of workplace based assessment[14]

- The difference between plan B and real improvement through innovation

If we see education also from the perspective of a business, it is worthwhile to make a distinction between the organisation's value proposition and the organisation's processes. As a result of the covert 19 pandemic, many educational organisations – including Australian colleges – have focused on adapting their current processes to an online-only context. In the short term, this has created some breathing space. There is another significant benefit from this application of the proverbial plan B, namely that it has 'loosened the existing processes sufficiently to enable true innovation. The medical education literature is now being populated with publications that describe experiences with moving

---

[10] - Prentice S, Benson J, Schuwirth L, Kirkpatrick El. A meta-analysis and qualitative analysis of flagging and exam performance in general practice training. *AUSTRALIAN JOURNAL OF PRIMARY HEALTH* 2019;25(3):XLIII-XLIII.

[11] - Govaerts MJB, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMMI. Workplace-Based Assessment: Effects of Rater Expertise. *Advances in health sciences education* 2011;16(2):151-65.

[12] - Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. *Advances in Health Sciences Education* 2013;18(4):559-71.

[13] - Ten Cate Th J. Entrustability of professional activities and competency-based training. *Medical Education* 2005;39:1176-7. doi: 10.1111/j.1365-2929.2005.02341.x

[14] - Valentine N, Wignes J, Benson J, Clota S, Schuwirth LW. Entrustable professional activities for workplace assessment of general practice trainees. Medical Journal of Australia. 2019 May;210(8):354-9.
- Weller JM, Misur M, Nicolson S, Morris J, Ure S, Crossley J, Jolly B. Can I leave the theatre? A key to more reliable workplace-based assessment. British journal of anaesthesia. 2014 Jun 1;112(6):1083-91.

processes online and lessons that can be drawn from that.[15] In addition, there are publications emerging which advocate for educational organisations to consider more revolutionary changes to their business.[16] There is now a unique opportunity to align educational processes with the imperatives of competency-based education, to extend the assessment tool box from a purely measurement orientation to one that also includes human judgement and due process, and finally, to smooth and the transition between the various phases of the education continuum from the first day of the undergraduate curriculum to the a final day of continuing medical education. Another reason to consider these fundamental changes exists because of the fundamental changes in the learners' affordances. Especially through ICT, learners now have affordances that did not exist in the past[17]; not in the least the continual availability of information everywhere through the Internet. Educational programs that do not sufficiently adapt to these fundamental changes and keep on thinking in terms of tweaking existing processes rather than a fundamental reorientation of their value proposition, run the risk of making themselves vulnerable. So, for organisations whose role is to ensure quality of health professions workforce in a country it is an important consideration whether they want to exert this role purely from a gatekeeper perspective or from the perspective of promoting of quality of all learners. The former typically leads to testing, whereas the latter would lead to a more longitudinal assessment program intertwined with feedback and educational activities.

In summary, for any redesign of assessment, especially within an academic/scientific context, there is consolidated evidence in the medical education literature from which appropriate strategies can be drawn. Unfortunately, a lot of that evidence is not in complete alignment with current practice and tradition. Approaches we believe to be valid and reliable have repeatedly been demonstrated to be all but valid and reliable. It is not an easy task to change assessment approaches in an existing organisation[18], but given the pandemic, the vulnerabilities of the existing (business) models and the rapid improvements and innovations across the globe, there is a real need and opportunity for a fundamental redesign of assessment practices.

[15] Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. Medical teacher 2018;40(12):1208-13.

[16] Hauer KE, Lockspeiser TM, Chen HC. The COVID-19 Pandemic as an Imperative to Advance Medical Student Assessment: 3 Areas for Change. Academic Medicine 2020

[17] Friedman LW, Friedman HH. The new media technologies: Overview and research framework. Available at SSRN 1116771 2008

[18] - Harrison CJ, Könings KD, Schuwirth LW, Wass V, van der Vleuten CP. Changing the culture of assessment: the dominance of the summative assessment paradigm. BMC medical education. 2017 Dec;17(1):1-4.