Review of the Alternative Assessment Pathway of the Royal Australian and New Zealand College of Psychiatrists. (27 November 2022)

Professor Lambert Schuwirth

Contents

Executive summary

The review of the Alternative Assessment Pathway (AAP) was commissioned by the RANZCP to consider the validity of the AAP and provide the College with advice and direction that will inform the future program of assessments.

The AAP was implemented initially as an emergency solution in lieu of the OSCE following the AV OSCE failure in November 2021 to accommodate eligible candidates' need for urgent assessment and progress them to Fellowship.

The AAP consisted of the three most recent In Training Assessments (ITAs) and a portfolio review. The ITAs are informed by Work Based Assessment (WBA)activities using Entrustable Professional Activities (EPA)s.

This report provides an evaluation of the quality and validity of the AAP and its process as an equivalent and alternative pathway to the OSCEs, and one that provided an assessment platform that supported candidates in the COVID environment throughout 2022.

In summary, the AAP was a longitudinal, multi-instrument, multi-supervisor assessment program in the authentic and complex practice and it replaced a purported single-event, single-instrument, centrally administered high-stakes assessment.

The combination of the WBA-ITAs and the Portfolio Review processes are better aligned with modern views on the assessment literature and the nature of competence being acquired in Psychiatry training. The use of WBA tools and EPAs outcomes which are carefully mapped onto competency domains and with good support for the narratives that can be used in the process demonstrate as definite strengths of the AAP.

Based on the process and tools used for the AAP, and the current understanding and insights of medical education and assessment research, the AAP has been therefore a better and more defensible assessment than the OSCE.

But this longitudinal component of the process only works if all stakeholders are committed to and engaged in the process, and the RANZCP supervisors are sufficiently 'assessment literate' to enable to contribute valid judgements on the trainees' performance portfolios.

The portfolio judgement process was careful and respectful and was seen as fair and holistic according to the current literature insights on modern assessment theory. The members of the Portfolio Review Oversight Panel were well trained for their task and engaged in the process with comprehensive level of fairness, respect, and seriousness. This aspect was also a strength of the AAP.

However, the use of a Case-based Discussion (CbD) as a second component of the AAP in the event candidates had an unsatisfactory outcome of the ITA-portfolio review, does not show good educational merit and is not of better standard than the OSCE. Although the outcome of this process in terms of false-positive outcomes is no worse than with an OSCE, it would still need to be the most urgent next step in any quality improvement, and any future assessment program may require a redesign or reconsideration of the CbD if it is used at all in the redevelopment of assessment strategy.

Because the AAP relies on human judgements made by experts, the recommendations in this report focus on the need to strengthen this aspect through ensuring quality of and access to supervisor training, exchange of expertise through communities of practice, support through building appropriate infrastructure to support assessments and/or extra supervisory support/engagement (for instance College-led remedial coaches) where needed. The role of the CbDs as part of the future process may need to be reconsidered and examined in the longer term – as assessment ***for*** learning - and can be further strengthened.

Provided the stakeholders are committed to genuine engagement in the future program of assessments and are sufficiently trained and supported, the AAP elements seem to be of at least equivalent quality as the OSCE and can actually be seen as an improvement and aligning to the contemporary assessment philosophy in medical education.

The RANZCP would be commended for taking further steps to transition its assessment into a workplace-based environment to provide a more flexible, contemporary, observable, and feedback-rich training and assessment that fosters learning consistency and progressive assessment of clinical skills and competencies that support the community needs of our two countries.

1 Introduction

Reviews in education or of an assessment program are never objective. They are typically written from a specific perspective and a specific interpretation of the relevant literature. These perspectives determine what aspects will be covered in the review and how they will be evaluated. That way it is clear how they have formed the basis for recommendations. To enable the College to appreciate the content of this review and the recommendations therein, I will first describe my perspective on assessment and its function, and then proceed to summarise what I think are the most important and most relevant lessons from the literature on assessment in medical education.

Assessment can be seen from various perspectives. One of the most popular perspectives on assessment is that it is a sort of prevention activity. From this perspective, the reason we have assessment is to prevent learners from becoming incompetent professionals, who would be likely to compromise patient safety. Hence, the focus of assessment is almost exclusively on identifying incompetent learners and holding them back until they can demonstrate sufficient competence or alternatively advising them to leave training. The assessment is then equivalent to a diagnostic test for an illness of 'incompetence'. In this 'prevention' metaphor, quality of assessment is defined as its ability to correctly discriminate between competent and incompetent students/trainees, or, conversely, to minimise false positive and false negative outcomes.

In this perspective, the impact assessment has on learning is mainly seen from a behaviourist approach. Passing the assessment and being allowed to progress to a next phase or to graduate can be seen as a reward or reinforcement, whereas failing the assessment and having to redo/repeat parts of the training program/education, or even having to leave or being excluded from the training program, is then seen as punishment for unsuccessful performance.

For a long time, this view has dominated the literature. Assessment was, therefore, mainly seen as a ***measurement*** of competence. From that point of view, high reliability and good construct validity, and objectivity of the process were seen as hallmarks of quality of assessment. The traditional OSCE was an instrument that was developed from this perspective on assessment, as an instrument to measure 'skills' or 'consultation ability'.

A popular view on competence at the time was based on a combination of knowledge, skills, problem-solving ability, and attitudes. From an assessment point of view, each of these four components was assumed to be a stable trait, like personality traits in psychology. With that view came the assumption that each trait could and should be measured reliably and validly at one point in time. This one-off assessment approach was seen as defensible because each trait was assumed to be stable. In addition, it was assumed that each trait could be measured independently of any other. Skills for example were supposed to be measurable independently of knowledge and vice versa. Such a view was epitomised in the early approaches to the assessment of skills with OSCEs[1]. The early versions typically consisted of very short – five-minute – stations that focused entirely on the technical skills without including any of the background knowledge. The early OSCE stations were very narrowly focused; a station on communication skills, for example, did not score for

[1] Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education* 1979;13(1):41-54.

relevance and content correctness of the communication, whereas more history focused stations did not include a score on the quality of the communication. Because 'skills' was assumed to be a stable and generic trait, the scores on all stations were simply added up to a total score.

In summary, this view stated that competence was a stable and generic phenomenon that could be objectively measured and should be expressed as a numerical outcome.

A considerable body of research emerged in the early days of assessment in medical education, and it led to some surprising and sometimes counterintuitive findings. Yet, these findings were quite robust.

1.1 Competence is domain specific

It is probably fair to say, that domain specificity is one of the most counterintuitive findings in the assessment literature. Domain specificity means that performance on one case, essay or OSCE station is an extremely poor predictor of performance on any other case, essay or OSCE station. For example, inter-station correlations are typically in the range of 0.1 or 0.2, or even negative. So, when a candidate performs very well on a knee examination it does not mean they will also perform well on an abdominal examination, or alternatively if they performed poorly on a neurology station that does not mean they will also perform poorly on a musculoskeletal station.

This finding, first described by Elstein, Shulman and Sprafka[2],was counterintuitive because up until then it was believed that competence was a generic quality. It was, for example, customary to have long case Vivas as single-event high-stakes examination. A popular notion at the time was that digging deep into a single case had higher validity and was more reliable than addressing multiple cases more superficially. Given the firm belief in this approach, it took quite a while before the issue of domain specificity was generally accepted and used in the design of more reliable (and therefore more valid) assessment programs. The most important implication of domain specificity is that assessment cannot be reliable if it is based on small samples and that for high-stakes assessments broad sampling is always needed[3]. This also means that standardisation or structuring the assessment has a limited positive effect on reliability, and that even with less standardised or structured approaches but with better sampling strategies higher reliabilities are achieved.

1.2 Medical problem-solving as idiosyncratic

Idiosyncrasy means that when different recognised experts are asked to solve the same problem, they are likely to employ different strategies or follow different solution pathways. This finding applies predominantly to assessment of clinical reasoning or clinical problem solving. There was more agreement from experts on some key decisions in the process, and generally they agreed on

---

[2] Elstein AS, Shulmann LS, Sprafka SA. Medical problem-solving: An analysis of clinical reasoning. Cambridge, MA: Harvard University Press 1978.

[3] 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.

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.

Norman GR, Van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education* 1991;25(2):119-26.

Van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education* 1991;25:110-8.

the outcome of the process, but the specific problem-solving pathways were quite idiosyncratic. In hindsight, this is logical because the way we solve problems, especially more complex problems, is based on our individual specific experiences. Our individual knowledge networks are what we use to determine the specific solution pathways. For the assessment of clinical reasoning or medical problem solving, this was an issue of concern, especially for those assessment methods that mimicked real life using long, branched simulations; the so-called patient management problems[4]. It turned out that these patient management problems seemed authentic and were therefore assumed to be valid, but because of the overriding issues of expertise idiosyncrasy and domain specificity they were neither reliable nor valid. In response to these concerns, an initial development focused on key decisions in the process where there was more consensus between the experts (key-feature approach or case-based assessment[5]) or on the outcome of the case (for example with extended matching items[6]).

But just focusing on the key decisions or the outcomes was not satisfactory for most assessment designers and a desire remained to assess the problem-solving or clinical reasoning process and not just the outcome. Most written types of assessment struggled with achieving this. For example, the script concordance test which aims to focus on decision nodes in the problem-solving process is not without controversy[7]. But also, for oral assessment types, including workplace-based assessment, the assessment of clinical reasoning or problem-solving is not without issues. An important new insight is that it should be perceived as a so-called complex adaptive process[8].

1.3 Validity

Validity pertains to whether the assessment assesses what it purports to assess. This is because we cannot directly observe competence. Competence exists in the heads of the learners, and we can only infer this from what the learner does or says. Whether that inference is correct or not, is the domain of validity. Traditionally, validity was seen purely from the perspective of predictive validity. A typical question would then be whether the test result predicts future performance as a doctor. But such predictive validity – or criterion validity – requires a gold standard. You can only determine predictive values if you have absolute agreement on the parameters of good clinical practice. In selecting such measurable outcomes for a gold standard, there will always be debate as to whether

---

[4] Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987;12(3):220 - 46.

[5] Bordage G. An alternative approach to PMP's: the "key-features" concept. In: Hart IR, Harden R, eds. Further developments in assessing clinical competence, Proceedings of the second Ottawa conference. Montreal.: Can-Heal Publications Inc 1987:59-75.

[6] Case SM, Swanson DB. Extended-matching items: a practical alternative to free response questions. *Teaching and Learning in Medicine* 1993;5(2):107 - 15.

[7] Charlin B, Brailovsky C, Leduc C, et al. The Diagnostic Script Questionnaire: A New Tool to Assess a Specific Dimension of Clinical Competence. *Advances in Health Science Education* 1998;3:51-8.

Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. *Medical Education* 2013;47(12):1175-83.

[8] Cristancho S, Field E, Lingard L. What is the state of complexity science in medical education research? *Medical Education* 2019;53(1):95-104. doi: https://doi.org/10.1111/medu.13651

the chosen gold standard is a valid predictor for good clinical practice[9] . As such, a view of validity which is purely based on criterion or predictive validity creates an ongoing problem of how to operationalise validity, ad infinitum. Instead, it was suggested by Cronbach and Meehl in 1955 to use a so-called construct validity perspective instead.[10] Construct validity does not try to predict future performance but employs a rigorous approach to determining whether the assessment actually measures the construct it purports to measure. So, if I were to design a new assessment which I claim assesses clinical reasoning, I would have to start with mounting a strong theoretical argument for the characteristics of the construct 'clinical reasoning' and then rigorously test whether my new assessment is valid for this construct. Since the work of Cronbach and Meehl, validity theory has undergone some changes, but fundamentally it is still based on mounting a series of arguments, including logic, plausible rationales, and testable research outcomes, to make the inference from observation to construct[11].

1.4 Reliability

Reliability is the extent to which outcomes on the assessment are generalisable. Every assessment is only a sample out of the whole domain of possible items or assignments. An OSCE for example, only contains a certain number (typically 10 to 16) of stations out of a large domain of possible stations that could have been chosen; even a 150-item multiple-choice test is only a sample of 150 questions out of an almost infinite domain of questions that could be asked. So, it is important that the result every candidate obtains is a good representation of that whole domain and that it is not just a matter of luck of the draw. Because it is impossible to assess that whole domain, reliability is typically defined as test-retest reliability. In other words, if the candidates were given a second test of equal difficulty and on the same topics but with slightly different questions, would they again obtain the same score, or the same position in the rank ordering from best performing to most poorly performing candidate, or at least, would they again pass or fail. However due to fatigue and the fact that students may have learned from the first test, administering a second test for the determination of test-retest reliability is not practically feasible. Various proxy approaches to a true test – retest reliability have been used.

The most intuitive one is probably the split half method, in which the test is randomly divided into halves and the correlation between one half and the other is used as a proxy for a test-retest reliability[12]. Because this randomly chosen split half can spuriously lead to a high or low correlation and may not be reflective of the true internal consistency, a more popular method is the average of all inter-item correlations, which is the basis of the most popular used reliability coefficient: Cronbach's alpha.

---

[9] Tamblyn R, Abrahamowicz M, Dauphinee D, et al. Effect of a community-oriented problem-based learning curriculum on quality of primary care delivered by graduates: historical cohort comparison study. *BMJ* 2005;331(7523):997-8.

[10] Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological bulletin* 1955;52(4):281 - 302.

[11] Kane M. Current concerns in validity theory. Journal of Educational Measurement. 2001;38(4):319-42.
Kane MT. Validation. In: Brennan RL, editor. Educational Measurement. 1. Westport: ACE/Praeger; 2006. p. 17 - 64.

[12] This would also require a correction for diminished test length

1.5 The role of content and assessment format for validity

Although generally, assessment methods are described by their format (OSCE, MCQ, SAQ, etc.) it is not the format of the assessment that determines its validity but the content. This is a robust finding in the literature. In other words, ***what*** the assessment asks is more important for its validity than ***how*** the responses of the candidates are recorded. It is a counter-intuitive finding, however, as we often assume that validity is associated with the format. A popular assumption, for example, is that open ended questions are better for the assessment of higher-order cognitive skills (such as clinical problem solving) than multiple choice. The research, however, consistently shows otherwise[13]. A striking example of this is a study which compared the results on an OSCE with those on a written (true-false item) test on clinical skills and showed that the correlations between both tests were extremely high[14]. The finding that the content is more important than the format is important in the consideration on how to combine results of various assessments in a program of assessment and forms a basis for the current views on programmatic assessment.

1.6 The role of human judgement in assessment

While for a very long time, objective and quantitative approaches to assessment were seen as the only pathway to fair and defensible outcomes and decisions, it became increasingly clear that there were aspects of competence that are not suitable for an objective and measurement-oriented approach. Communication, empathy, collaboration, health advocacy, etc. are more complex aspects of competence and require subjective human judgement rather than a measurement approach[15]. With respect to this development, it is important to understand that subjectivity is not the same as unreliability and objectivity is not the same as reliability[16]. But even subjective human judgement must not lead to unfairness in the process[17]. Several strategies can be used to achieve this. One is to ensure procedural fairness in the process, for example through transparency, audit trails and clear communication, but another important strategy is to ensure that examiners are sufficiently trained and have sufficient assessment expertise to produce a valid judgement. Modern assessment programs do not shy away from including human judgement as well as measurement-based assessment methods.

1.7 Assessment of learning versus assessment for learning

A popular current distinction is between assessment ***of*** learning and assessment ***for*** learning. Assessment of learning is the more traditional approach in which the purpose of the assessment process is to determine whether the learner has achieved sufficient competence or made sufficient progress. Traditionally, this is done with single-event high-stakes examinations, like many traditional approaches to college examinations. However, assessment ***of*** learning programs still has an impact

---

[13] Cf for a summary of the literature: Schuwirth LWT, Van der Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education 1996;30:44 - 9.

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

[15] Valentine N, Durnig SJ, Shanahan EM, et al. Fairness in Human Judgement in Assessment: A hermeneutic literature review and conceptual framework. *Advances in health sciences education* under editorial review

[16] Norman GR, Van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education* 1991;25(2):119-26.

[17] Valentine N, Durning SJ, Shanahan EM, et al. The pursuit of fairness in assessment: Looking beyond the objective. *Medical Teacher* 2022:1-7.

on the learning behaviour of the candidate. But this impact is predominantly through behaviourist drivers of punishment (failing) and reward (passing). In medical education, there is an increasing dissatisfaction with an exclusive assessment ***of*** learning approach. The main reason for the dissatisfaction is the so-called Goodhart's law. Goodhart's law states that as soon as a measure becomes a target it ceases to be a good measure. Or in other words, if an outcome becomes a target, people will try to achieve the outcome with whatever process is most effective and efficient. Translated to the assessment context, this means that when cheating or plagiarising is a more effective and efficient way of achieving a high grade than learning/studying, some learners will engage in cheating or plagiarising. In an assessment ***for*** learning approach, the focus is not on a one-time measurement of achievement but on using assessment to foster the learning process longitudinally. So, rather than focusing on the achievements or results of an assessment, the assessment program assesses the way the learner performs, receives, and engages with feedback, formulates learning goals and undertakes targeted learning activities.

On a subsequent assessment, the learner is then able to demonstrate that they have improved in the areas they were given feedback on. So, assessment ***for*** learning is not the same as formative assessment. Formative assessment focuses on providing the learner with feedback and leaving it up to the learner whether to use the feedback, assessment for learning requires the learner to engage meaningfully with the feedback and demonstrate the necessary improvement. Failure to engage or show the necessary improvement leads to summative consequences of failing the assessment.

1.8 Assessment as a program

Based on the robust finding that the content of the assessment is more important for its validity than the assessment format, programmatic assessment is an approach that starts with a principle that assessment information must be collated and synthesised through a meaningful narrative[18]. In the assessment context, this is still rather counterintuitive, but it mimics clinical practice very well. In clinical practice, it is not standard to add up information because it is of the same format. For example, one would not be inclined to disregard a high glucose level because of a low sodium level. In other words, clinicians are not inclined to use a compensatory approach on those two lab values simply because they are both lab values. Instead, the finding of a high glucose level will be narratively combined with the patient's complaints of thirst, fatigue and frequent urination and perhaps physical examination findings of poorly healed wounds and weak peripheral arterial pulsations. Programmatic assessment follows the same principle of combining parts or passages of various assessments to construct a meaningful narrative and understanding of the learner's performance, progress, and achievements.

1.9 Complexity and non-linearity

In the conceptualisation of the nature of competence, and subsequently the approaches to education and assessment, a major shift has taken place. Roughly until the 2000s, the dominant view saw competence as a straightforward phenomenon, a so-called latent trait that could be objectively measured. Education was also seen as a linear process; the teacher's role was predominantly to lecture and tell the students what they need to know, and the assessment was mainly focussed on the students demonstrating what they had learned or remembered from the teachings. This is, of course, a slightly exaggerated and caricatural description, but it illustrates the thinking about education and assessment of that time. But such a perspective does not really work

[18] Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Medical Education. 2005;39(3):309-17.

well in workplace learning and the workplace-based assessment context[19]. In this practical context, learning cannot be scaffolded or predicted but occurs based on 'what walks through the door'. Modern views on competence, and consequently, on education and assessment are, therefore, increasingly embracing complexity perspectives.

Complexity perspectives differ from linear views in many ways. One important difference, for instance, is that in linear views the ***components*** of the system are seen as essential, whereas in a complexity view, the ***interaction between*** the components is seen as more essential. So, a traditional OSCE applies a linear view when its focus is on whether all the steps (checklist items) have been completed. As a result, there is only one optimal solution to the case or station. More modern assessment approaches have moved away from this view and, using more subjective human judgement, acknowledge that the essence is not merely ticking all the items of a checklist but to form a judgement about how the candidates string the items together.

To use a clinical analogy, it is like the difference between a patient history as a set of survey questions or as a doctor-patient communication. This also means that more modern approaches to assessment acknowledge that there are multiple, equally good ways to solve a problem. The linear perspective can still be useful as an approach to the assessment of factual knowledge or even application of knowledge. However, more complex aspects of competence, such as communication, collaboration, cultural competence, require complexity views. Multiple and diverse observations and judgements in practice are then a better and more logical approach to assessment.

One other difference is important. A linear approach is more often associated with a so-called deficiency model, while complexity views embrace a so-called diversity model. By this, I mean that when an assessment leads to a (numerical) score, the underlying assumption is that that competence can be measured and expressed on a single scale. The corollary is that a candidate who achieves a lower score is more deficient in competence than a candidate who achieves a higher score. With complexity views and multiple solutions that can be seen as acceptable, this also means that differences between candidates can be accepted, and that diversity is acknowledged. Obviously, in linear views, quality is defined as standardisation of the process and equality, whereas in complexity views the focus lies on standardisation of the quality of the process and equity.

2 Used resources

For this report I have had the opportunity to interview key stakeholders in the assessment program of the College, who have given me access to various websites and other resources. The main resources used for this report are:

- https://www.ranzcp.org/pre-fellowship/about-the-training-program#Program%20structure
- https://www.ranzcp.org/pre-fellowship/assessments-workplace/wbas
- https://www.ranzcp.org/files/prefellowship/2012-fellowship-program/epa-forms/epa-table.aspx

- https://www.ranzcp.org/files/prefellowship/2012-fellowship-program/ranzcp-ita-stage-1-end-rotation.aspx

[19] Actually, it is also increasingly seen as an outdated perspective in the more theoretical phases of medical education.

- https://www.ranzcp.org/pre-fellowship/about-the-training-program/stage-1#Stage1EPAs

- https://www.ranzcp.org/epahandbook.aspx

- https://www.ranzcp.org/files/prefellowship/2012-fellowship-program/administration/learning-outcomes.aspx

- https://www.ranzcp.org/files/prefellowship/2012-fellowship-program/administration/developmental-descriptors.aspx

- Example of all forms, mini-CEX, presentation, Case-based discussion and direct observation of procedural skills and the ITA form.

- During the meeting I have seen examples of the Excel portfolio summaries

3 Overview of the assessment program

The College employs a longitudinal workplace-based assessment program with regular decision moments. The whole program takes a minimum of 60 months of training subdivided into periods of six-months each. Stage 1 takes 12 months and stage 2 and 3 each take 24 months. The main approach to assessment is by using Entrustable Professional Activities (EPAs)[20]. There is a prescribed sequence in which the EPAs are completed during the training program. A minimum of 16 EPAs must be completed by the end of training stage 2 (36 months) and the additional EPAs depend on the particular choice of elective rotations. In total, most trainees have completed between 25 and 30 EPAs at the end of their training.

The decision to use EPAs as outcomes is defensible given the existing literature. EPAs provide a way of defining outcomes that is more in line with the day-to-day vocabulary of clinical supervisors and they align more with the types of decisions about entrustment that supervisors are used to making routinely. The use of these entrustment decisions rather than a more education-oriented vocabulary has been shown to improve the reliability and validity of workplace-based assessment judgements[21].

Theoretically, there would be three possible ways of using EPAs across the whole training continuum.

The first would be to use the same EPAs from start to finish and at every assessment point, evaluate the trainee's progress towards the endpoint of full entrustment in all EPAs. This, however, presupposes that the trainee is given an opportunity to develop in all domains of the EPAs, or, in other words, have exposure to all kinds of patients, subdisciplines, et cetera.

The second would be to require completion of EPAs during training but allow optimal flexibility with respect to which EPA is going to be completed at which specific point in time during the training.

[20] Ten Cate O. Entrustability of professional activities and competency-based training. Medical Education. 2005;39:1176-7.

[21] Weller JM, Misur M, Nicolson S, Morris J, Ure S, J C, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. British Journal of Anaesthesia. 2014;112(6):1083-91.

That way, there is more flexibility in matching the assessment requirements with the affordances of practice or 'what walks through the door'. Such flexibility would be useful because workplace-based learning or workplace-based education is less structural and planned than, for example, preclinical training. However, it does require the supervisor, who signs off the certificate of entrustment possesses, to have sufficient expertise in the domain of each EPA to warrant a valid assessment.

The third approach is the one currently undertaken by the College with a prescribed sequence of core EPAs and some flexibility in the rest, depending on the specific elective rotations chosen by the trainee. Given the current workplace organisational structure, this third approach is currently most defensible and seems to be the most plausible in the use of EPAs in the RANZCP's program. However, this does not preclude any future consideration of changing the regulatory structure of the assessment program to allow greater flexibility as to which EPA can be granted at any point in the program. Such a consideration must be weighted of course against feasibility in the program, buy-in from stakeholders/supervisors and directors of training, and any unwanted strategic assessment behaviour of trainees.

The content of the EPAs is carefully mapped against the end-of-training learning outcomes. These learning outcomes are based on the CanMeds framework and consist of the competency domains of medical expert, communicator, collaborator, manager, health advocate, scholar and professional. Each of these domains has subdomains which are extensively described in the learning outcomes. Such a description is important as it makes the expectations for each stage of training concrete. They provide the 'vocabulary' – the 'signs and symptoms of competence - for supervisors to make their decision[22]. More importantly, the EPAs are developed as they are mapped onto the learning outcomes. For example, EPA **ST1-GEN-EPA5** maps onto the subdomains 1,2,3,4,5 of the learning outcome medical expert, subdomains 1,2 of the learning outcome communicator, subdomains 1,2,3 of the learning outcome collaborator, et cetera.

The decision whether to award a trainee an EPA with a certificate of entrustment (COE), is informed by the more formal WBA instruments (mini-CEX, professional presentation, Case-based discussion, direct observation of procedural skills and OCA) in conjunction with general observations of the trainee's performance and information from others. The OCA stands out from the rest currently as it is a required activity but not necessarily an assessment that has to be passed.

Thus, the assessment program is based on longitudinal observations with more formal assessment moments over a range of practical contexts. Only supervisors will undertake formal assessment activities with the registrars, however informal observations and feedback from other stakeholders can also be used as information sources. Using EPAs, these informal observations and formal assessment are collected, collated, and eventually synthesised in the in-training assessment (ITA).

In each stage of training, two ITAs are conducted. The first, generally mid-stage, is formative in nature and the second towards the end of the stage is summative.

The collection and collation of all relevant information for the final assessment in the ITA is supported by an electronic learning management system, which is used by almost all supervisors.

[22] Valentine N, Schuwirth L. Identifying the narrative used by educators in articulating judgement of performance. Perspectives on medical education. 2019;8(2):83-9.

The whole process is predominantly qualitative, and the forms used in the assessment – either the WBA assessments or the ITA – use ordinal and descriptive criteria – and no numerical scores are used in the practical assessment program. From a traditional psychometric perspective this may raise concerns, but from the perspective of the modern assessment literature this is a strength.

Modern views on assessment and competence hold that competence is not seen as a realist phenomenon that can be objectively measured but is only meaningful through rich and descriptive narrative. In other words, providing feedback to a learner and educational organisation that decides about the learner's performance is better if that feedback contains rich and meaningful information ***for*** learning. Extensive narrative/feedback collected and collated over a longer period and synthesised in an expert fashion whilst ensuring transparency, accountability and fitness for purpose is currently considered as the most defensible and fair approach to assessment.[23]

It is important to note that using a WBA approach as an OSCE alternative entails a move towards a more modern approach to assessment and is conceptually a step forward.

A corollary of this modern view on competence and assessment programs is that the validity and reliability – fairness – are not exclusively a feature of the assessment program. It is not just in the assessment methods and the rules and regulations, but also in the users.

For such a program to remain strong, it is paramount that all users, or at least all supervisors and directors of training, have sufficient knowledge about the program, how to use it and how to make defensible judgements of performance and competence. This is often referred to assessment literacy[24]. Without such assessment literacy, judgements of competence and performance are likely to be less valid. Therefore, attention to a well-designed assessor training program is advisable. How to manage such a training program in the context of an organisation like the College is not straightforward. On the one hand, would centralisation of such training, i.e., keeping everything under control of the College, allow for control of quality, but it would also run the risk that individual directors of training and supervisors feel a loss of ownership over their own expertise development. While centralised processes may lead to equality – everyone receives the same treatment – this can come at the expense of equity, everyone receiving the training that works best for them.

Hence, while the idea of a central training program and centrally administered examinations may seem appealing, the literature argues these may not be the best models to approach medical education. A delegated structure has advantages of being more locally relevant and applicable and provides a healthy feeling of ownership. The risk is that it might lead to unwanted variability of training quality. If there is an indication there is undesirable variability of training quality for existing and new supervisors or directors of training, the following steps may be considered:

1 Invest in production of generic training tools (videos, role play scenarios, background documentation, etc) which can be adapted and used to suit the local training program.

[23] Valentine, Nyoli, et al. "Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement." *Medical Education* 55.9 (2021): 1056-1066.

[24] Popham, W. James. "Assessment literacy for teachers: Faddish or fundamental?." Theory into practice 48.1 (2009): 4-11.

Berendonk, Christoph, Renée E. Stalmeijer, and Lambert WT Schuwirth. "Expertise in performance assessment: assessors' perspectives." Advances in Health Sciences Education 18.4 (2013): 559-571.

2 Establish and support a platform in which individual training programs can exchange experiences, learn from each other, and seek to further improve their training practices – as a community of practice.

Training is not only important to ensure general validity and reliability of the assessments, but also to prevent leniency bias. Having to tell a trainee that their progress or achievement is unsatisfactory and that they are not ready to progress to a next stage is always an uncomfortable conversation. But it is an important one, as the literature shows that leniency bias typically sets up the learner for failure in the future[25].

3.1 In-Training Assessments (ITA)

ITAs take place during every rotation. In the rotation there are typically two ITAs, one mid-rotation and one at the end of the rotation. The mid-rotation ITA is formative and is used to discuss with the trainee strength and weaknesses and targeted areas for improvement. The end of rotation ITA has a more summative character and is used to determine whether the trainee is ready to progress to a next rotation. ITA forms have been adapted to the specific phase of training and whether they it concerns mid-or end of rotation ITAs.

Retrospective final judgements can be vulnerable in the assessment program when they are either based on the general, retrospective impression and/or on second-hand information. This, however, does not appear to be the case with the in-training assessments of the College. The ITA combines information from workplace-based assessments, general observations, and the EPAs.

In an ideal situation, this information is collected and collated in the electronic learning management system of the College. This means that information is firsthand, recorded and documented at the time of observation, and certified by documentation such as certificates of entrustment. A second strong component of the ITA process is the mid-rotation ITA. This step ensures that the trainee is aware of their strengths and weaknesses and is provided with feedback and areas that require improvement. Theoretically, the rotation outcome should not come as a surprise to the trainee.

However, in practice there may be issues that interfere with this process. For example, a serious issue emerges in the final weeks of training which has consequences for the summative outcome or where there have been pre-existing issues that may not have been sufficiently identified and/or documented or may have been observed towards the end of the rotation only. Ideally these situations are rare, and struggling trainees are identified and managed as early as possible.

Based on literature, this combination of feedback, agency of the learner and the final outcome which does not present a surprise to the trainee is seen as a token of quality of the assessment of a program. Furthermore, this system enables the supervisor to initiate the formative ITA process at an earlier point in time during the rotation, if the supervisor thinks the registrar is struggling. In a consensus paper on quality design of assessment programs several tips are suggested, such as:

- develop an example process that promotes a feedback orientation
- adopt a robust system for collecting information
- make sure that low stakes assessment provides meaningful feedback
- provide mentoring to learners
- ensure trustworthy decision-making
- promote continuous interaction between stakeholders

[25] Papadakis, Maxine A., et al. "Disciplinary action by medical boards and prior behavior in medical school." *New England Journal of Medicine* 353.25 (2005): 2673-2682.

All these aspects are, in principle, built into the ITA program with the possibility for the supervisor to engage the Director of Training and, in partnership with the trainee, develop a support and remediation plan. It is the combined responsibility of the supervisors' and the trainees' that the remediation plan and follow up are executed effectively.

Focusing on remediation has two key reasons.

The first is that a combination of identifying struggling trainees and targeted remediation has been demonstrated to lead to better outcomes as the end[26] result.

The second is that failure to identify issues early and having a leniency bias or trusting that they will be dealt with at a later stage is not fair to the trainee and sets them up for bigger failures in the future[27]. Once a trainee enters a remediation process - with a minimum of three months – they should be able to demonstrate sufficient remediation/improvement through the normal ITA process. There is a maximum of 3 failed rotations, after which the trainee would have to show cause or undertake a Training Review (RANZCP). This means that remediation has a maximum duration of 18 months. Although there might be circumstances beyond the trainee's control that would warrant longer remediation, it is best to cap the process duration, or else it would theoretically carry a risk trainee might take up many 'remediation' opportunities until they may eventually pass - maybe by chance. That could be seen as a false positive pass.

Two considerations are important though. The first is that the quality of the whole process depends on the quality of the human factor in the process. Any form of workplace-based assessments that requires observations and judgements is dependent on the assessment expertise, motivation, and engagement of its stakeholders. The second is the quality of the contributing information on progress, specifically the documented observations of workplace-based assessments.

The role of the assessment-specific expertise of the supervisor has already been discussed above. The next section will consider the specific WBA tools.

3.2 Workplace based assessments

3.2.1 Mini-CEX

The College's mini CEX procedure adheres to the same principle as originally described in the paper by Norcini et al.[28] But there are some interesting differences. The original mini CEX was clearly intended to serve purely as an assessment of learning/summative assessment component. The College approach differs in several aspects.

From the website information, it reads that the supervisor randomly selects the patient based on availability or that a new patient is used in the mini CEX. However, there might be circumstances in which the supervisor feels that it is better to purposively select a patient for the min-CEX. Alternatively, there can be trainees who are doing very well and who are an 'agentic' partner in the selection of a patient. From an assessment ***for*** learning perspective and in the context of adult

[26] Prentice, Shaun, et al. "Identifying the at-risk general practice trainee: a retrospective cohort meta-analysis of general practice registrar flagging." *Advances in Health Sciences Education* 26.3 (2021): 1001-1025.
[27] Papadakis, Maxine A., et al. "Disciplinary action by medical boards and prior behavior in medical school." New England Journal of Medicine 353.25 (2005): 2673-2682.
[28] Norcini J, Blank LL, Arnold GK, Kimball HR. The Mini-CEX (Clinical Evaluation Exercise); A Preliminary Investigation. Annals of Internal Medicine. 1995;123(10):795-9.

learning in the training program, this is a highly defensible approach. However, keeping the ultimate responsibility with the supervisor lends credibility to the process from an assessment ***of*** learning perspective. The latter is important because every workplace-based assessment gives input into the EPAs and through that into the ITAs and final summative decision. Thus, it is important that the supervisor can assume responsibility and accountability for the quality of this assessment component with respect to determining whether the trainee is progressing well enough.

The form is well designed. It starts with ample space for feedback before the assessment criteria are scored. The assessment criteria are not scored numerically but purely as ordinal values. This is a second aspect in which the college's mini-CEX differs from the original description. The 'scoring' is simply an indication whether the candidate progresses as expected or not. The form contributes to the identification of a trainee who may need remediation.

Additionally, the form is designed to allow/show a complexity view. Different trainees can achieve the same ordinal outcome but with different strengths and weaknesses as a diversity perspective, while a numerical scale (such as the typical 9-points scale) would suggest a deficiency model. It is important to note that in workplace-based assessment, the validity of the assessment relies heavily on the interaction between the form and the user[29]. In this case, the form appears to be optimally designed to be used by an expert and experienced clinical supervisor. In summary, the combination of the guidelines and instructions that are available online, the process of selecting patients for the activity and the form design give the impression of being designed in a thoughtful and evidence-based manner.

3.2.2 Professional presentation

The WBA professional presentation is an assessment that is focused more on communicating knowledge, understanding or management of a patient. This assessment relies on mutual trust and respect between supervisor and trainee in relation to choosing a topic, evaluating performance, providing feedback, and deriving possible learning goals and activities from it. The form is well designed, not overly prescriptive, and therefore well-suited for the specific purpose.

3.2.3 Direct observation of procedural skills (DOPS)

The direct observation of procedural skills is a WBA assessment is aligned and similar to the mini CEX but is more focused on the practical/procedural aspects of consultation. The characteristics of the WBA component are similar to those mentioned for the mini CEX.

3.2.4 Case-based discussion,

The Case-based Discussion differs from the other WBA components in that the trainee has more agency with respect to the choice of cases. The trainee submits four cases of which the supervisor will select one. In the context of training and as one of the components of a more extensive program of assessment this is the defensible and for most of the trainees probably a valuable approach. From the perspective of assessment in the context of an adult learning pathway with assessment ***for*** learning, there is obvious evidence of strength in the process of making choices of assessment material in consensus between the supervisor and the trainee through mutual trust and respect. However, some external stakeholders might be inclined to still view such an assessment purely from an assessment ***of*** learning perspective and as a measurement of competence.

[29] Govaerts, M. J. B., et al. "Workplace-based assessment: raters' performance theories and constructs." *Advances in Health Sciences Education* 18.3 (2013): 375-396.

3.2.5 Observed Clinical Activity

The observed clinical activity (OCA) is another workplace-based assessment component, with a slightly different status in the WBA program. It is a compulsory activity that must be completed but does not require successful performance. Yet, failure to complete an OCA would lead to a 'fail' decision on the ITA, and it is therefore a barrier requirement. In the WBA program, the OCA can potentially play a confusing role for the trainees. It is an activity that must be completed but does not indicate the required standard level. It is likely that trainees assume that the performance on the OCA will play a role in the ITA considerations by the supervisor, but this is not formally articulated.

Such dual messaging can be potentially confusing. What is also slightly confusing is how the OCA form differs from all the other WBA forms. These WBA forms use a three-point scale which is ordinal and asks for a judgement on whether the progress and achievement of the trainee is at the expected level given their stage in training, below the expected level or above the expected level. The OCA form however employs the traditional nine-point scale. I don't think that the nine-point scale adds much to the three-point scale, any nuances are better recorded in the open space for feedback.

3.2.6 Summary of the WBA program

In summary, the five WBA components appear to be valid and suitable for the purpose. But it needs to be reiterated that the quality (validity, reliability, and fairness) of any observation based or workplace-based assessment depends not only on the rules, regulations, and the forms, but how they are used by supervisor and trainee. The overall design of the WBA program, the EPAs and ITA and the way they are combined is a definite strength from an assessment design perspective. The many direct observations with mini CEX, OCAs, CbD, professional presentations and DOPS, with a clear mapping on the learning outcomes and EPAs makes for a very coherent and well-defined assessment program. However, there is some complexity which requires supervisors, trainees, and directors of training to be well informed and have a good understanding of how the program works.

Any lack of understanding or misconceptions could compromise/threaten the quality of the assessment process. Hence, the assessment expertise and/or literacy are paramount – this is especially pertinent as no specific educational or assessment training - other than being a fellow - is required to be eligible for the role of supervisors. The learning management system, InTrain, can play an important role in supporting stakeholders with collecting, documenting, collating and eventually synthesising all information in the expected way.

In addition to the components that have been discussed, there are also developmental descriptors which may not be extensively used. It is difficult to understand how and where they would be used and what added value they offer to an already well described program of assessment.

As stated previously, subject to the process being followed correctly, all information – from informal observations and formal assessment activities – is collected and collated in a (electronic) dossier and that all those involved have been provided sufficient training in assessment to acquire sufficient assessment literacy, I would state that the WBA as part of the AAP is a more defensible assessment approach than the OSCE[30].

However, it is important to be mindful there might be external stakeholders who hold a view on assessment that it is still based on an assessment ***of*** learning and from a perspective of central control. They may not be inclined to see the strengths of the current WBA program. It is, therefore, important that the College has a strong and convincing narrative about strengths and weaknesses of

[30] Of course the WBA program is only part of the AAP

the WBA component of the AAP but also about the strengths and weaknesses of the OSCE. I hope that the introduction of this report is useful in constructing or further strengthening this narrative.

3.3 Portfolio review

The WBA program as described in the previous sections is only a component of the AAP. The final decision whether the performance of the trainee is comparable to passing an OSCE examination is made during the portfolio review. The requirement is to have completed at least 30 months of training and at least three of the most recent ITA results. If these contain a stage 3 ITA with at least 3 months FTE duration, a decision is made based on this portfolio. If that trainee has not completed at least 3 months FTE in stage 3, they will be required to sit a Case-based discussion.[31]

Portfolios are reviewed by portfolio review panels. These review committees are provided with a summary overview of the information contained in the ITAs for decision-making. The process is 'fire-walled'; the members of the portfolio review committee are not involved in the direct supervision of training of the candidates whose portfolios they are reviewing. The portfolio summaries are completely de-identified but if a panel member suspects they recognise a trainee from a pattern in the portfolio they will have to declare a conflict of interest and excuse themselves and the portfolio is taken to another panel.

In most cases, consensus can be reached about candidates' outcomes. When the panel members disagree about the outcome, the portfolio is referred to the Portfolio Review Oversight Panel (PROP) with the deliberations of the original panel.

The panel strives to ensure that each decision is optimally fair for every trainee. The literature describes various steps that can be incorporated in such a process to ensure fairness[32]. One of these components is benchmarking, which in this case means undertaking a calibration process. Where most of the candidates' outcomes might be clear passes, and a small minority may be clear fails, there will be results in the 'grey area'. Ideally, members of portfolio review panels would have had some sort of calibration session – similar to the CBD calibration activity – to ensure that they can reach a defensible decision for trainees who are in this grey zone. This is not to ensure a more or less 'true' decision, but to ensure that the decision will be seen as fit for purpose, transparent, credible, and defensible. An alternative to a calibration process, especially if it would only pertain to a minority of the trainees, would be cross consultation between review panels. In the case of the AAP the PROP plays an important role of a strong 'second pair of eyes' or a mediator.

In conclusion, the portfolio review process is defensible and fair because:

- it is based on longitudinal information as recorded in the ITAs, therefore over multiple occasions, with multiple instruments and generally by multiple observers.
- it is based on prolonged engagement between supervisors and trainees and the more formal WBA activities and therefore not dependent on snapshot judgements of 'independent' examiners at one point in time, like in the OSCE

[31] Separate rules apply to Specialist International Graduates; their portfolio review will involve the three most recent ITAs, and those who have completed fewer than three will be considered on a case-by-case basis

[32] Valentine, Nyoli, et al. "Fairness in human judgement in assessment: a hermeneutic literature review and conceptual framework." *Advances in Health Sciences Education* 26.2 (2021): 713-738.

Driessen, Erik, et al. "The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study." *Medical education* 39.2 (2005): 214-220.

- a careful two-person review of the summary of the information by the portfolio review panel
- a second look by the four-person PROP.

3.4 Case-based Discussion

If a trainee's results are deemed not to be sufficiently meeting the required standard in the Portfolio Review, a Case-based Discussion (CbD) takes place. Candidates are asked to prepare two clinical cases in which they have been involved and for which they have secured patient's consent. The two assessor Panel will then select one of these cases for the presentation by the candidates as CbD assessment. The assessors may ask additional or follow-up questions to explore the breadth of the case in relation to learning outcomes. The whole process takes one hour and 15 minutes, with 15 minutes pre-reading and preparation, 45 minutes of actual CbD and 15 minutes for the assessors to complete the marking. The assessment is conducted online, either via Zoom or with Teams as backup.

There may be two reasons for a candidate to proceed to a CbD. The first is that the outcome of the portfolio review may be borderline. This means that no clear decision could be made as to whether the candidate was successful. In this case, the CbD can be seen as an additional source of information as a sequential assessment approach.

Another reason for a candidate to proceed to the CbD is if the portfolio review resulted in an 'unsuccessful' outcome and the candidate has not demonstrated the required standard. In that case, the CbD is not an additional source of information or sequential assessment, but a resit or repeat assessment which can possibly overturn the outcome of whole the portfolio review process.

From a theoretical perspective it is less defensible that an examination based on a single case, even with a duration of 45 minutes, can overturn an assessment made based on a whole portfolio containing three ITAs.

I think there are huge strengths to the WBA/EPA process in conjunction with ITAs and the careful portfolio review process. These strengths provide clear narrative and connections between multiple observations in practice and assessments during and at the end of a rotation. A further strength is the 'longitudinality' of collection and collation of information, the combination of intermediate assessments with feedback and access to remediation, and well-informed summative progress decisions.

The program includes opportunities to identify struggling trainees and offer early remediation, and, finally, the close collaboration between supervisor and trainee in managing the whole assessment process can add considerable strengths to the assessment program and learning process. This allows for convincing evidence of candidate competence than an OSCE as a single high-stake assessment.

However, the ITA/portfolio review process is also considerably more convincing than a single CBD. As mentioned earlier in the introduction, domain specificity is an overriding issue of concern in any form of assessment. While domain specificity is unlikely to play a role in the longitudinal workplace based/ITA program, it is very likely to play a role in a single CbD.

This is compounded by the fact that the trainees can select and submit their own cases. From an assessment ***for*** learning perspective this is, as I explained earlier, desirable, but at this point in the process with the high-stakes decision to be made, an assessment ***of*** learning perspective takes precedence.

From a societal point of view, false-positive results (candidates passing who are not sufficiently competent yet) would probably be a greater concern than false-negative outcomes. For the November 2021 AAP cohort, a total of 15 candidates who were considered 'unsuccessful' passed the CbD out of a total of 184 who were eligible for the AAP, which is 8.15%. For the March 2022 AAP cohort, 17 of the 'unsuccessful' candidates' out of 324 eligible candidates passed the AAP, which is 5.25%. To put these numbers into a perspective, it would be meaningful to compare them to general OSCE outcomes. For an OSCE with a reliability of .75 the upper 95% confidence interval is roughly 1 standard deviation wide[33]. Assuming a normal distribution, the area under the curve between the cut-off score and the upper 95% confidence interval would be .3413. This means a 34.13% likelihood of a result being a false-positive (or not meeting the p≤ 0.05 requirement) if the cut-off score would be close to the mean of the distribution. It would be .1159 (11.59%) if the cut-off score were 10% higher than the mean.

However, score distributions on OSCEs are generally not normally distributed but left skewed, in which case the percentages (areas under the curve) become even bigger. So, although the CbD as a resit for the portfolio can be seen as the weaker aspects of the whole AAP, the results of this process (8.15% and 5.25%) and the likelihood of false-positive results as a consequence do not appear to be worse than would occur with a standard OSCE process.

While the literature suggests that in such types of assessment, resources are better spent on having one examiner per case but multiple cases, it is understandable that if the decision is made to only have one case, examiners feel safer in examining in pairs.

Hence, it might be advisable to consider having a CBD process in which both submitted cases are examined for 20 minutes, but with one examiner for each case. This would constitute the same time and resource investment but would likely lead to better reliability[34].

Finally, consensus marking does not add to the validity of reliability of the process. Different examiners have different views/perspectives on the competence of a candidate. These perspectives are most likely to be complementary to each other[35]. By requiring a consensus marking, the assessment information from these different perspectives is lost. Instead, it would be better to retain the marking from each individual assessor and decide afterwards whether the candidate can be granted fellowship or not. If insufficient agreement can be reached, it is more logical to have an additional CbD than to try to reach consensus.

However, it is logical that any assessment process needs to have a follow-up for unsuccessful candidates, and this might have been the reason for the case-based discussion component. It is absolutely an understandable design decision, but it would be one that I would see as the first target for further improvement of the whole program and future assessment direction.

[33] The standard error of measurement in classical test theory is $SEM = SD \times \sqrt{(1-\alpha)}$, and with α = .75 (and the square root of .25 being .5), the SEM is half a standard deviation. The 95% confidence interval is 1.96 x the SEM and therefore, with α =.75, the 95% confidence interval is roughly +/- 1 standard deviation (.98 standard deviation to be exact).

[34] with only one case and consensus marking, reliability cannot be calculated for this approach to CbD

[35] Gingerich, Andrea, et al. "Seeing the 'black box' differently: assessor cognition from three research perspectives." Medical Education 48.11 (2014): 1055-1068.

4 Conclusion and recommendations

4.1 Conclusion

In conclusion, the Alternative Assessment Pathway (AAP) has considerable strengths and if used well I would see it as an improvement over the single-event high-stakes OSCE. It is longitudinal over a longer period enabling broad sampling, using multiple instruments as formal assessment components, and it combines the observations and judgement of a range of supervisors, formal assessment with more informal observations, provides a system for discussing the outcomes with trainees, and for the provision of ample feedback. It also facilitates collection and documentation of available assessment and performance information and has a coherent – narrative – system for translating this information into EPAs as per the CanMeds-based learning outcomes.

Another reason why I see the AAP as an improvement over a single-event high-stakes examination (like an OSCE) is because it better aligns with the complexity of some competencies, such as communication, collaboration, advocacy, and most importantly cultural competence – often misnamed 'soft skills'. These are aspects of competence that cannot be sufficiently tested with an assessment approach that adheres to a linear view (such as an OSCE or any other single-event high-stakes examination) but requires assessment that embraces complexity, which is not only important for the validity of the assessment of the 'soft skills', but also because it enables a more inclusive and diverse approach to trainees, especially those from diverse cultural backgrounds, and most certainly with respect to First Nations trainees - a diversity approach is an improvement over a deficiency perspective.

Provided all stakeholders, especially supervisors and Directors of Training, are fully informed and on board with the purpose and intended processes of the assessment and are well experienced and have the required expertise to make valid judgements during the WBA activities and beyond, it is a robust, trustworthy, and credible program, and highly defensible. The only area I have some concerns about is the use of case-based discussion as a resit for a failed portfolio review (or even as an additional source of information).

For a sequential approach, extended assessments which are information richer would be more suitable (multiple CbDs and not selected by the candidate). For those candidates who are 'unsuccessful', a further directed rotation based on the identified weaknesses would be a more defensible alternative approach, although this may not be practically feasible (at this point in time).

It is important to explicitly state that reviewing an assessment program from interviews and documentation will never include a deep understanding of and appreciation for the possibilities and limitations in the organisation. Organisations have their specific culture which strongly impacts the level of acceptability of an assessment program and the willingness of stakeholders to seriously engage with it. This willingness to engage and acceptability are essential to the quality – validity – of a workplace-based assessment program. Therefore, any recommendations in this report must be seen as suggestions for targeted areas for improvement, and it is important that decision-makers and educational and assessment designers in the college are able to gauge the optimal timelines for any changes. COVID has been a disruptive time, and it has forced many organisations to quickly change assessment practices and adopt opportunities for change and improvement. Some of these changes have been successful and some may have experienced less fortunate implementations. The disruption of the last two years has also placed huge demands on College Fellows, trainees and SIMGs and that many are feeling fatigued. In this context I would offer the following recommendations.

4.2 Recommendations

- The College consider developing a suite of centrally designed tools or programs for induction and training of supervisors and directors of training. This would enable each training region to use and adapt these tools to their own specific context. In addition, the college may consider offering a platform to develop a Community of Practice through which different training regions can exchange and share experiences about their training or can even consult each other about difficult training situations.

- The College may consider designing a program to support the requirements of the new training processes for supervisors and Directors of Training who are undertaking additional roles as assessors in the new assessment program, including further upskilling of existing supervisors and training as a form of 'accreditation'. It may include peer support and consultation, calibration, development of consistent assessment rubrics and tools.

- All Directors of Training and supervisors should be enabled to use the electronic learning management system to ensure that all trainees are assessed using the same infrastructure. This can help with ensuring equity across all trainees and training regions.

- The College may consider ways to provide additional support to supervisors and/or Directors of Training with struggling trainees. It takes less time, energy, and resources to pass a trainee than to hold them back and provide remediation, and this can make a WBA system vulnerable as there is a bonus to leniency bias or mercy passes. Extra support could counteract this. Counteracting leniency bias is important because it leads to an increase of problems and is likely to set up the trainee for failure in the future.

    - Part of this support may be the development of a range of online resources such as training modules for supervisors in relation to difficult conversations or how to deal with the trainee in difficulty. Another consideration might be for the college to have access to special remedial supervisors/coaches. In organisations where such special remedial coaches are employed (such as GP training), the experiences have been positive.

- The College should disseminate a clear narrative about the background and rationale for giving the trainee agency in the process. From an assessment ***for*** learning and adult learning perspective and given the fact that there is a clear relationship between supervisors and trainees, in conjunction with the careful documentation and transparency of the whole process, this trainee agency is defensible. However, external stakeholders coming from a different assessment perspective may need convincing. In my opinion, the College would benefit from having clear communication materials to convince such external stakeholders, and the literature overview in the first section of this report may be helpful.

- The College should develop a strong and succinct narrative or communication to explain the role of complexity and that the elements from the AAP are more conducive to a diversity model than a single-event high-stakes examination (which is more based on a deficiency model). This narrative/communication should highlight both the more plausible validity for the complex – 'soft'- competencies and how it caters better to a diversity of trainee backgrounds and a more inclusive training program. The College should provide clear rationale for any program that it develops and seek consultation / collaboration on decision making.

- There is a clear intent in the whole program to be development oriented. This is an absolute strength given the context of the RANZCP training. From modern views on assessment, we have come to understand that when assessment focusses on development of the learner, with a longitudinal, multi-perspective, multi-instrument assessment program is likely to produce higher levels of competence than a purely measurement-oriented assessment of learning[36]. The process could be further strengthened if it included a more formal connection between assessment moments.

    - The feedback and outcomes of one assessment moment (for instance one mini-CEX) will have to be the starting point for the next. That way, the feedback loop is mandatorily closed. An analogy can help explain how this will work. When a researcher submits a manuscript to a journal, it is peer-reviewed and they are given the decision 'major revisions', there is the expectation that the researcher will create a table of reviewer and editor feedback in which the researcher describes precisely how they have responded to this feedback and then submit a revised version of the manuscript with track changes.

    - In a true assessment ***for*** learning program, feedback is given during and at the end of an assessment moment, and it is the expectation that the learner will accept and record this feedback, design, and initiate learning activities to address it, and is able to demonstrate that they have successfully actioned the feedback on the next assessment occasion. Since all the assessment moments are already in place in the current program with the multiple WBAs (including OCAs) and general observations, building in a standard process of reviewing past feedback and learning activities is not a major step, but one that has significant impact on the ability of the program to be a developmental assessment for learning.

- The College may consider replacing the case-based discussion as the resit assessment with a more information-rich approach (and may also consider this for those cases in which the CbD serves as an additional/sequential source of information). This would be recommended as a first step in the future development of assessments.

    - In the shorter term, the College may consider changing the case-based discussion process to include more cases (more than one for assessment). For example, it could consider requiring all trainees to submit four cases from which 2 will be selected. Each of those two cases can then be examined by one examiner and the final decision is based on the outcome of both cases. That way the sample of cases is increased (from 1 to 2) and so is the sample of examiners (from a consensus mark to two independent scores). This would address two major issues, domain specificity, and less undesirable control (with opportunities for strategic assessment behaviour) for what is basically an assessment ***of*** learning event.

    - Another minor recommendation is regarding the role of the developmental descriptors in the whole process, to provide better clarity on how and where they should be applied and what they add to the existing rich narrative of the descriptors in the process. I recommend the College consider either making their role clearer in the whole program or abandoning them.

[36] Heeneman, Sylvia, et al. "Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment." *Medical teacher* 39.1 (2017): 44-52.