



# Refresh of a Clinical Skills Assessment for Physician Trainees

Elizabeth Whiting <sup>1-3</sup>, A Curtis Lee <sup>4</sup>, Balakrishnan R Nair <sup>5,6</sup>

<sup>1</sup>Cross College Examination Review Advisory Group, Royal Australasian College of Physicians, Sydney, NSW, Australia; <sup>2</sup>Faculty of Medicine, University of Queensland, Brisbane, Qld, Australia; <sup>3</sup>Internal Medicine Services, The Prince Charles Hospital, Brisbane, Qld, Australia; <sup>4</sup>School of Medicine, University of Notre Dame, Sydney, NSW, Australia; <sup>5</sup>School of Medicine, University of Newcastle, Newcastle, NSW, Australia; <sup>6</sup>Centre for Medical Professional Development, Hunter New England Local Health District, Newcastle, NSW, Australia

Correspondence: Elizabeth Whiting, Internal Medicine Services, The Prince Charles Hospital, Rode Road, Chermside, Brisbane, Queensland, 4032, Australia, Tel +61 419730908, Email [elizabeth.whiting@health.qld.gov.au](mailto:elizabeth.whiting@health.qld.gov.au)

**Purpose:** The Royal Australasian College of Physicians (RACP) oversees physician training across Australia and Aotearoa New Zealand. Success in a written examination and clinical skills assessment (known as the clinical examination) at the mid-point of training is a requirement to progress from basic to advanced training. The clinical examination had evolved over many years without a review process. This paper describes the approach taken, the changes made and the evaluation undertaken as part of a formal review.

**Methods:** A working party that included education experts and examiners experienced in the assessment of clinical skills was established. The purpose of the clinical examination and competencies being assessed were clarified and were linked to learning objectives. Significant changes to the marking and scoring approaches resulted in a more holistic approach to the assessment of candidate performance with greater transparency of standards. Evaluation over a 2-year period was undertaken before the adoption of the new approach in 2019.

**Results:** In 2017 testing of a new marking rubric occurred during the annual examination cycle which confirmed feasibility and acceptability. The following year an extensive trial utilising the new marking rubric and a new scoring approach took place involving 1142 examiners, 880 candidates and 5280 scoresheets which led to some minor modifications to the scoring system. The final marking and scoring approaches resulted in unchanged pass rates and improved inter-rater reliability. Feedback from examiners confirmed that the new marking and scoring approaches were easier to use and enabled better feedback on performance for candidates.

**Conclusion:** The refresh of the RACP clinical examination has resulted in an assessment that has clarity of purpose, is linked to learning objectives, has greater transparency of expected standards, has improved inter-rater reliability, is well accepted by examiners and enables feedback on examination performance to candidates.

**Keywords:** physician assessment, reliability, clinical assessment, long case, short case

## Introduction

Physician training in Adult Medicine and Paediatric and Child Health in Australia and Aotearoa New Zealand is overseen by the Royal Australasian College of Physicians (RACP). The RACP is in line with health professional education programs nationally and internationally and is adopting a programmatic assessment (PA)<sup>1</sup> approach. Experience has shown that successful implementation of PA in medical schools requires intensive work including renewal of curricula, fundamental changes in assessment design and intensive support for learners and teachers.<sup>2,3</sup> Adopting PA in the RACP and ensuring standardisation is additionally challenging given the significant number of trainees and supervisors distributed across many hospitals in two countries. Moreover, the implementation timeframe will, by necessity, be protracted. In the context of challenges related to implementation of PA in the RACP and the likely lengthy timeframe required, a decision was made that at least in the medium term, to continue the traditional RACP assessment approach including written and clinical examinations. In 2015, a review of the clinical examination was initiated to ensure it constructively aligns with PA during its implementation.

## Background

### Physician Training Program

The RACP physician training program takes a minimum of 6 years. Trainees must satisfactorily complete all training requirements related to the 3 years of basic training before being eligible to progress to the written examination and assessment of clinical skills (known as the clinical examination). Success in the examinations is followed by at least 3 years of advanced training in a chosen sub-specialty. Satisfactory completion of all advanced training requirements is followed by admission to Fellowship of the RACP (FRACP). Approximately 1000–1200 trainees participate in the annual clinical adult and paediatric examinations across Australia and Aotearoa New Zealand.

### Clinical Examination

#### Format

The format of the clinical examination has evolved over many years and includes two long and four short cases over a day of examination. Both long and short cases utilise real patients and are conducted face to face in a hospital setting. Long cases are comprised of an unobserved one-hour consultation with a patient followed by a 25-minute discussion with the examiners. Short cases consist of an observed physical examination of the patient and discussion with the examiners over a 15-minute period. Cases are selected to assess a diversity of medical sub-specialties. Candidates are examined outside of their own hospital and geographical region.

#### Examiners

Two examiners are required for each of the six cases in the examination. One of the examiners is typically a physician from the hosting hospital and the other an experienced physician from a national panel. Of the pair, one examiner will be the lead examiner for each case examined. On the day of the examination but prior to its commencement, each examining pair reviews the cases they will examine on. This is conducted in circumstances matching those of the candidates with a similar time allocation and without access to clinical notes or other pre-emptive information. The examiners agree on issues to be addressed, clinical signs and expectations for a pass standard. Over the course of the examination day, a candidate will be examined by eight physicians. In many of the examining teams there may also be an observing physician who is preparing to become an examiner in the following year. Examiners are drawn from all sub-specialties, although examiners do not participate as the lead examiner if the case is primarily related to their sub-specialty. For each case, examiners mark separately and then determine a final score based on discussion and consensus. Each year a small number of national examiners volunteer to participate in calibration and examinations in both countries.

#### Calibration

The purpose of calibration sessions is to familiarise examiners with the assessment rubrics and the exam process and to establish the passing standard. Both local and national panel examiners participate in annual mandatory calibration sessions prior to the examination. Calibration includes instructions on the conduct of the exam with opportunities for discussion. On an annual basis, prior to calibration, a short case and a long case calibration video is made using real patients. These videos are viewed by the entire National Examiner Panel at their annual calibration and individually marked according to the standardised RACP marking system. This is followed by discussion and agreement on a consensus final score. These videos and the consensus scores are then utilised the following year to guide calibration at the local calibration sessions.

#### Scoring

Until 2018 the scoring system was based on a 19-point scale. Long cases were weighted three times the value of short cases. The sum of the scores determined the overall mark with a candidate requiring a minimum of 120 points out of a total of 210 possible points to pass the examination. An additional requirement was that a minimum of a pass in one short case and a pass in one long case was required to achieve an overall pass.

## Methods

A working group was established which included education specialists, experienced RACP examiners and representatives of the examination and education committees. A project plan was approved including clarification of project scope,

a benchmarking and literature review, followed by an external expert and stakeholder consultation. The necessary criteria for good assessment are well described and include validity, reliability, equivalence, feasibility, educational effect, catalytic effect and acceptability.<sup>4</sup> The Van der Vleuten utility index includes five of these elements and is a conceptual model of assessment utility as key elements in the design of assessments.<sup>5</sup> Within the scope of the project, these elements were considered and provided a guide for the working party in the review.

In addition, issues identified by trainees and examiners over the years also needed to be considered. These issues included:

- The purpose of the clinical examination was unclear including competences being assessed.
- The competencies were not mapped to the basic training curriculum.
- The 19-point scoring system was complex for examiners to use and had suboptimal inter-rater reliability.
- Giving candidates useful feedback was challenging in the context of a 19-point scoring system.
- The cumulative scoring system was potentially overly simplistic and did not give opportunities for a holistic assessment.
- Examination outcomes were heavily affected by the weighting of the long case scores.

## Review

### Purpose Clarification

Purpose clarification was an essential first step. The goals of the examination were reviewed and updated and were then linked to the learning and assessment objectives (Boxes 1 and 2). All subsequent innovations were designed to be consistent with the purpose of assessment.

### Format

With an updated clarity of purpose, consideration of approaches as outlined in national and international literature and feedback from internal and external consultation, a decision was made that the traditional format of the examination would continue unchanged. The working party affirmed the validity of the examination based on assessment of real patients in hospital settings across a range of skills integral to physician practice. Examination by calibrated examiners

#### Box 1 Statement of Purpose of the Clinical Examinations

The purpose of the RACP Divisional Clinical Examination is to assess the clinical skills, clinical acumen and interpersonal skills to inform whether trainees have reached the standard for completion of Basic Physician Training

#### Box 2 Purpose of the Long and Short Case

The purpose of the **LONG CASE** is to test clinical assessment skills through discussion with emphasis on:

- Accuracy of the history,
- Accuracy of the clinical examination,
- Synthesis and prioritization of clinical problems,
- Understanding the impact of the illness on the patient and family,
- Development and discussion of an appropriate management plan.

The purpose of the **SHORT CASE** is to test clinical assessment skills through direct observation and discussion with an emphasis on:

- Interaction with patient/family,
- Examination technique,
- Examination accuracy,
- Interpretation and synthesis of physical findings,
- Investigations/management.

who had undertaken blinded patient assessments prior to the candidate assessment and the inclusion of two long cases and four short cases enabled standardisation and reliability of results.

## Marking Rubric

The scoring system was simplified from the 19-point scale to a 6-point scale. Criteria for assessment of performance for both long and short cases were updated and aligned to the domains as outlined in the purpose statements. The domains were also mapped to the basic training curriculum. A descriptor for each score across the domains was developed to ensure clarity of the expected standard and summarized in marking rubrics for the long (Figure 1) and the short (Figure 2) case. Examiners use the assessment domains to guide them in arriving at an overall final mark depending on the emphasis of the domains in the individual case being examined.

In 2017, the new marking rubrics were piloted to assess acceptability and feasibility of the new approach. One pair of senior examiners at each examination site was asked to use both marking systems for each candidate with only the traditional marking approach being used to determine the candidate's final result. Examiner feedback confirmed that the new marking rubric was preferred to the traditional approach with a small number of minor alterations.

## Scoring Grid

There were anecdotal concerns related to the triple weighting of long case scores. It was noted that simply summing results across cases and the triple weighting of long cases in comparison to short cases could lead to results that may not reflect the overall performance of the candidate. A strong performance in one long case despite relatively poor performances in the other cases could result in a candidate passing. Similarly, a poor performance in one long case,

ADULT MEDICINE		LONG CASE		CRITERIA FOR ASSESSMENT OF PERFORMANCE			
FACE - TO - FACE INTERVIEW VERSION							
ASSESSMENT DOMAINS >		ACCURACY OF HISTORY	ACCURACY OF THE CLINICAL EXAMINATION	SYNTHESIS & PRIORITISATION OF CLINICAL PROBLEMS	UNDERSTANDING THE IMPACT OF THE ILLNESS ON THE PATIENT AND FAMILY	DEVELOPMENT AND DISCUSSION OF AN APPROPRIATE MANAGEMENT PLAN	
LEVEL OF PERFORMANCE	6	<b>Excellent Performance</b>	<ul style="list-style-type: none"><li>Sophisticated interpretation of the history</li><li>Focuses on key issues</li><li>Shows perceptiveness in extracting difficult information</li><li>No need to clarify details</li></ul>	<ul style="list-style-type: none"><li>Actively seeks subtle signs that might enhance diagnosis</li><li>Superior organisation of difficult examination</li></ul>	<ul style="list-style-type: none"><li>Identifies all major and minor problems</li><li>Very careful prioritisation which includes a long term view</li><li>Recognises social impact of disease</li></ul>	<ul style="list-style-type: none"><li>Shows mature understanding of subtle, difficult, or intimate aspects of patient's functioning</li><li>Demonstrates balance when discussing issues and sophisticated use of external social support</li></ul>	<ul style="list-style-type: none"><li>Superior construction of management plan, including long term impact</li><li>Highly developed and discriminating use of investigations</li><li>Mature recognition and interpretation of inconsistent results</li></ul>
	5	<b>Better than Expected Standard</b>	<ul style="list-style-type: none"><li>Emphasis on appropriate details</li><li>Appreciates subtleties</li><li>Interprets significant aspects of the history</li></ul>	<ul style="list-style-type: none"><li>Includes important relative negative signs</li><li>Appreciates significance of more subtle signs</li></ul>	<ul style="list-style-type: none"><li>Confidently identifies essential problems</li><li>Shows maturity in recognising lesser issues</li></ul>	<ul style="list-style-type: none"><li>Shows persistence in exploring subtle psychological issues, or issues that impact on the patient or family</li></ul>	<ul style="list-style-type: none"><li>Proposes appropriate management plan with good understanding of social impact lifestyle and psychological aspects of disease</li><li>Good use of discriminating investigations</li><li>Accurate interpretation of results</li></ul>
	4	<b>Expected Standard</b>	<ul style="list-style-type: none"><li>Reasonably complete, accurate and detailed history</li><li>Minimal need to clarify details</li><li>Timely and well structured</li><li>Some interpretation</li></ul>	<ul style="list-style-type: none"><li>Correctly identifies most important physical signs</li></ul>	<ul style="list-style-type: none"><li>Identifies all key problems</li><li>Arranges problems in order of priority</li></ul>	<ul style="list-style-type: none"><li>Understands patient's physical and psychological functioning in relation to disease</li><li>Appreciates impact of treatment and prognosis on patient and family</li></ul>	<ul style="list-style-type: none"><li>Proposes an appropriate and realistic management plan for the major issues</li><li>Provides a sensible, balanced approach to investigations</li><li>Interprets most investigations appropriately</li><li>Recognises important side effects of proposed treatment</li></ul>
	3	<b>Below Expected Standard</b>	<ul style="list-style-type: none"><li>Incomplete, inadequately detailed and/or inaccurate history, and/or poorly timed</li><li>Need to clarify important details</li></ul>	<ul style="list-style-type: none"><li>Omission and/or incorrect reporting of some important signs</li></ul>	<ul style="list-style-type: none"><li>Problems poorly prioritised</li><li>Significant problems undervalued</li></ul>	<ul style="list-style-type: none"><li>Fails to recognise some important aspects of the disease on patient or family</li><li>Misses some aspects affecting functioning or reaction to illness</li></ul>	<ul style="list-style-type: none"><li>Some errors in arranging a management plan</li><li>Erratic and non-discriminatory use of investigations</li><li>Errors in the interpretation of tests</li><li>Lacking adequate appreciation of complications of treatment</li></ul>
	2	<b>Well Below Expected Standard</b>	<ul style="list-style-type: none"><li>Poorly organised</li><li>Omission of many key points</li><li>Inaccuracies or lack of detail</li><li>Repetitive, poorly structured</li><li>Historical details not clarified</li></ul>	<ul style="list-style-type: none"><li>Many significant signs not recognised</li></ul>	<ul style="list-style-type: none"><li>Poor understanding of significant problems</li><li>Requires substantial prompting</li></ul>	<ul style="list-style-type: none"><li>Poor understanding of the impact of disease on patient and family</li><li>Shows little concern about psychological aspects</li></ul>	<ul style="list-style-type: none"><li>Inappropriate or poorly directed management plan</li><li>Poor understanding of useful investigations</li><li>Inability to interpret investigations</li><li>Major inability to appreciate side effects of treatment</li></ul>
1	<b>Very Poor Performance</b>	<ul style="list-style-type: none"><li>No clear structure</li><li>Focused only on single problem</li><li>Minimal detail</li></ul>	<ul style="list-style-type: none"><li>Minimal attention to detail with the examination</li></ul>	<ul style="list-style-type: none"><li>Most key management issues unidentified</li><li>No attempt to establish priority</li></ul>	<ul style="list-style-type: none"><li>Impact of disease not explored at all, or unable to be discussed</li></ul>	<ul style="list-style-type: none"><li>Poorly directed management plan without consideration of major issues</li><li>Very poor ordering of investigations without consideration of expense or potential complications</li><li>No attempt to interpret investigations</li><li>No understanding of side effects of treatment</li></ul>	
EPA Competencies		EPA 1, EPA2 Medical expertise, communication, (cultural competence)	EPA 1 Medical expertise	EPA 1 Medical expertise, judgement and decision making	EPA 1, EPA2 Medical expertise, communication, ethics and professional behaviour, judgement and decision making, (cultural competence)	EPA 1, EPA4, EPA 6 Medical expertise, communication, ethics and professional behaviour, judgement and decision making	

NOTE: In coming to an overall assessment score, not all domains will be equally weighted or always applicable due to variability of patient cases

Version 1.7\_2 Face to Face • May 2021  
(identical to version 1.5 • March 2020)

NOTE: In coming to an overall assessment score, not all domains will be equally weighted or always applicable due to variability of patient cases

Version 1.7\_2 Face to Face • May 2021  
(identical to version 1.5 • March 2020)

**Figure 1** Long Case Rubric. EPAs (Entrustable Professional Activities) form part of the Basic Training Curricula Standards and refer to “essential work tasks trainees need to gain competence in, perform safely, and be entrusted by their supervisors to do in the workplace” (<https://www.racp.edu.au/trainees/basic-training/curricula-renewal/standards>, accessed 30/05/2024).



## ADULT MEDICINE

## SHORT CASE

## CRITERIA FOR ASSESSMENT OF PERFORMANCE



ASSESSMENT DOMAINS >		INTERACTION WITH PATIENT/FAMILY Candidates SHOULD achieve the expected standard in terms of their interaction with the patient/family	EXAMINATION TECHNIQUE	EXAMINATION ACCURACY	INTERPRETATION AND SYNTHESIS OF PHYSICAL FINDINGS	INVESTIGATIONS/ MANAGEMENT	
LEVEL OF PERFORMANCE	6	Excellent Performance	<ul style="list-style-type: none"><li>Exceeds expected standard</li></ul>	<ul style="list-style-type: none"><li>Highly fluent, accurate and within time</li><li>Makes adjustment to routine where appropriate</li><li>Includes and completes additional complementary examination elements unprompted</li></ul>	<ul style="list-style-type: none"><li>Correctly identifies all essential and desirable signs</li></ul>	<ul style="list-style-type: none"><li>Establishes the most likely diagnosis on the basis of examination</li><li>Provides a reasonable differential diagnosis based on physical findings</li><li>Considers all likely alternatives with a higher level justification</li><li>Able to rule out unlikely diagnoses</li></ul>	<ul style="list-style-type: none"><li>Correctly interprets investigations and integrates with examination findings without prompting</li><li>Recognises and discusses areas of doubt</li><li>Uses results to support differential diagnosis and discussion</li></ul>
	5	Better than Expected Standard	<ul style="list-style-type: none"><li>Meets expected standard</li></ul>	<ul style="list-style-type: none"><li>Fluent and accurate and within time</li><li>Makes adjustment to routine where appropriate</li></ul>	<ul style="list-style-type: none"><li>Correctly identifies all essential and most desirable signs</li></ul>	<ul style="list-style-type: none"><li>Identifies the most likely diagnosis</li><li>Provides a reasonable differential diagnosis based on physical findings</li><li>Considers likely alternatives with justification</li></ul>	<ul style="list-style-type: none"><li>Correctly interprets all major investigations</li></ul>
	4	Expected Standard	<ul style="list-style-type: none"><li>Introduces him/herself to the patient</li><li>Shows respect for patient as indicated by preservation of patient's modesty, seeking permission for sensitive aspects of examination</li><li>Recognises and modifies examination when painful</li></ul>	<ul style="list-style-type: none"><li>Undertakes systematic examination of required area or system without unnecessary duplication</li><li>Demonstrates confidence in the examination</li><li>Completes assigned tasks in appropriate time</li></ul>	<ul style="list-style-type: none"><li>Detects most essential signs</li><li>Reports significant negative findings</li><li>Does not find major signs that are not present</li></ul>	<ul style="list-style-type: none"><li>Provides sensible provisional diagnosis and discusses appropriate differential diagnoses</li><li>Recognises most inconsistencies in interpretation and findings</li><li>Provides sensible priorities in diagnosis</li><li>Does not propose diagnoses inconsistent with signs</li></ul>	<ul style="list-style-type: none"><li>Reasonable interpretation of investigations</li><li>Suggests appropriate line of investigation and integrates them with examination findings</li></ul>
	3	Below Expected Standard	<ul style="list-style-type: none"><li>Inappropriate and insensitive approach to patient</li></ul>	<ul style="list-style-type: none"><li>Examination incomplete or lacking fluency or systematic approach</li><li>Includes unnecessary duplication</li></ul>	<ul style="list-style-type: none"><li>Misses essential signs</li><li>Fails to look for or identify important negative findings</li></ul>	<ul style="list-style-type: none"><li>Not confident with a diagnosis and/or provides diagnoses not consistent with signs</li><li>List of differential diagnoses poorly developed and/or inconsistent with signs</li><li>Unable to consider alternative explanations for findings</li><li>Requires more than minor prompting to reconsider options</li></ul>	<ul style="list-style-type: none"><li>Does not offer appropriate investigations</li><li>Misinterprets or is unable to integrate investigations with examination findings</li></ul>
	2	Well Below Expected Standard	<ul style="list-style-type: none"><li>Unduly rough, clumsy or causes pain without adjustment or apology</li></ul>	<ul style="list-style-type: none"><li>Very slow and requires substantial prompting and guidance</li><li>Required examination incomplete</li></ul>	<ul style="list-style-type: none"><li>Misses essential signs</li><li>Finds abnormalities that are not present</li><li>Fails to look for important negative findings</li></ul>	<ul style="list-style-type: none"><li>Unable to suggest a reasonable diagnosis</li><li>Advances diagnoses inconsistent with signs</li><li>Requires substantial prompting</li><li>Unable to reconsider additional information which may alter diagnosis</li></ul>	<ul style="list-style-type: none"><li>Unable to use investigations to assist in diagnosis</li><li>Inappropriate dependence on investigations</li></ul>
	1	Very Poor Performance	<ul style="list-style-type: none"><li>Requiring examiners to intervene</li></ul>	<ul style="list-style-type: none"><li>Slow examination not completed in appropriate time</li><li>Cannot perform appropriate examination of system</li></ul>	<ul style="list-style-type: none"><li>Misses all essential signs</li><li>Finds abnormalities that are not present</li><li>Fails to look for important negative findings</li></ul>	<ul style="list-style-type: none"><li>Unable to suggest a reasonable diagnosis</li><li>Unable to interpret the physical signs elicited</li></ul>	<ul style="list-style-type: none"><li>Unable to suggest reasonable investigations</li><li>Misinterprets information provided</li></ul>
EPA Competencies		EPA 1, EPA 2 Medical expertise, communication, ethics and professional behaviour	EPA 1 Medical expertise, judgement and decision making	EPA 1 Medical expertise, judgement and decision making	EPA 1 Medical expertise, judgement and decision making	EPA 1, EPA 6 Medical expertise, judgement and decision making	

NOTE: In coming to an overall assessment score, not all domains will be equally weighted or always applicable due to variability of patient cases

Version 1.6 • April 2020

**Figure 2** Short Case Rubric. EPAs (Entrustable Professional Activities) form part of the Basic Training Curricula Standards and refer to “essential work tasks trainees need to gain competence in, perform safely, and be entrusted by their supervisors to do in the workplace” (<https://www.racp.edu.au/trainees/basic-training/curricula-renewal/standards>, accessed 30/05/2024).

even with satisfactory performances in most of the other cases would still likely result in the candidate failing. This was confirmed in modelling and identified the opportunity to test a new approach and refresh how an overall outcome might be determined.

The working group accepted that a significant change in expected pass rates for the examination was not in scope for the review, meaning that historical pass rates could be used to establish a baseline for the expected pass rates for the new scoring system. Simulations of retrospective data from 2015 to 2017 examinations were applied to several scoring grid models. After review, the working group agreed that the clinical examination should be considered as one examination that consisted of a combination of both long and short cases. Rather than numerical weighting of individual cases of the examination leading to the biases described above, a more holistic approach that takes performance in all six cases into consideration with an opportunity for compensation if a candidate does better in some cases compared to others was considered to be a preferential approach. However, sufficient differences between competencies tested by the long and short cases would also mean that a minimum standard would need to be achieved in both types of cases. The scoring grid needed to reflect this “compensatory” and “minimum standard” concept, address the problem of undue influence of performance in one case and achieve a comparable pass rate to the traditional approach.

In the development of the scoring grid a “Policy” approach<sup>6</sup> was adopted to set the passing criteria based on the level of performance. This entailed the creation and agreement on a set of rules to determine outcomes which were based on the following requirements:

- A minimal level of performance on each type of case was identified.

- A compensatory approach was required balancing performance on each type of case.
- Strong performance on one case would compensate for poorer performance on the others.

Initially, a grid was developed considering just the scores from the long cases. Candidates scoring poorly on both long cases would clearly be unsuccessful irrespective of performance in the short cases and those scoring highly on both would likely be successful although would need to achieve a minimum standard in the short cases. Given this, the preliminary grid was developed (Figure 3).

With the “compensatory” approach, each short case score contributes to the overall score, but the overall score required to pass would be determined by the strength of performance in the long cases. This was captured on the grid so that an increasingly strong performance in the long cases required progressively less strong performances on the short cases. A diagonal series of bands that outlined different groupings of performance was added where it was felt that the overall performance outcome would be considered similar.

Finally, rules outlining the minimum overall scores to achieve a pass standard for each band needed to be established. In order to determine the band rules, the number of short cases passed and the aggregate short case scores were considered in the context of performance in the long cases (Figure 4).

With this “compensatory grid”, examination outcomes for the new approach were able to be simulated and compared to the traditional approach with a focus on borderline candidates and those that passed with one approach and failed in the other. Upon review of the data, the working group was more confident that the results for these individuals utilising the new marking rubric was more reflective of overall performance in comparison to the traditional approach.

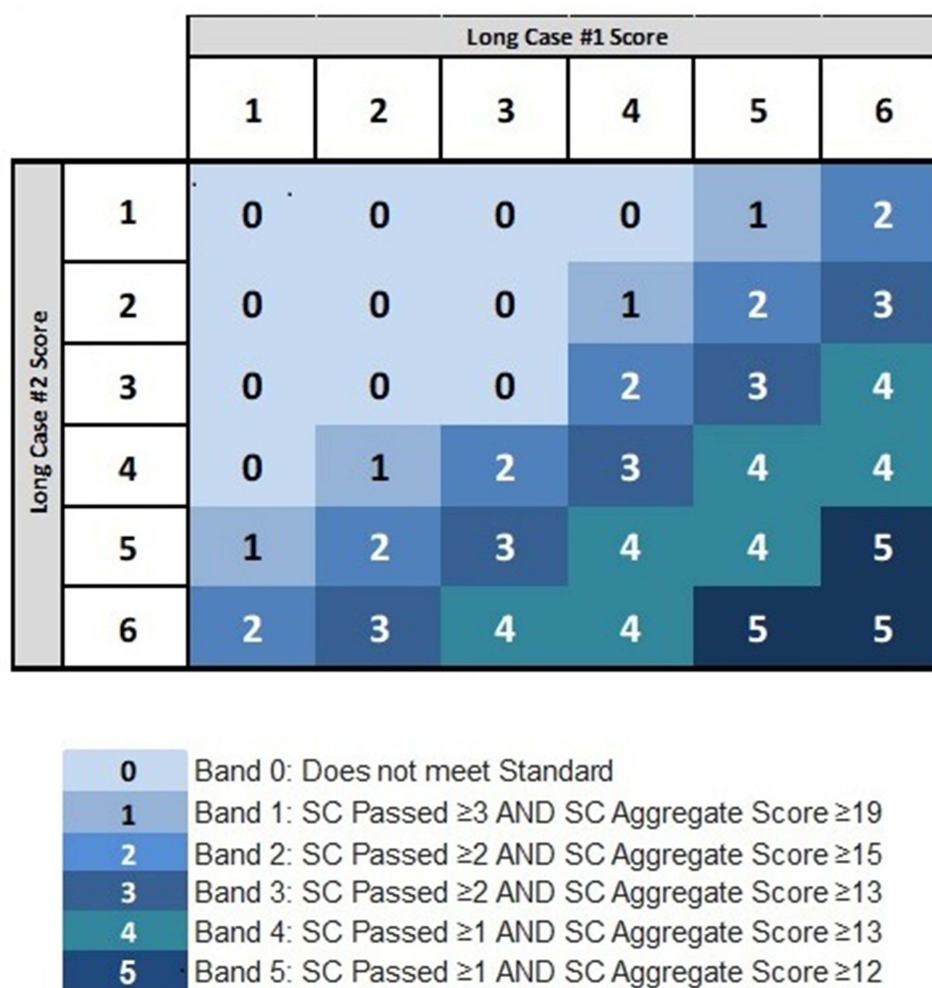
## Evaluation

As a high stake assessment, the project plan needed to ensure that the proposed changes were sound and that the psychometric performance of the instrument was acceptable. In 2018, an extensive trial took place involving 1142 examiners, 880 candidates and a total of 5280 scoresheets for individual cases from across Australian and New Zealand from both Adult Medicine and Paediatric and Child Health Divisions. Both the new approach and the traditional approach were used during the examination with only the traditional approach being used to determine candidate outcomes. This trial enabled further feedback from examiners on the marking rubric and provided paired data to enable further refinement of the scoring grid.

There was a 43.3% response rate (494/1142 examiners) in a post-examination examiner survey, which confirmed that the new marking rubric was operationally easier, enabled agreement on consensus scores and that there was overall support for replacing the traditional approach with a new rubric.

		Long Case #1 Score					
		1	2	3	4	5	6
Long Case #2 Score	1	Below Standard					
	2						
	3						
	4						
	5						
	6						Above Standard

**Figure 3** Combining Long Case Scores.



**Figure 4** Score Combination Grid.

**Abbreviation:** SC, Short Case.

The scores in the new and traditional approaches utilised in 2018 were examined in detail. The pass rate utilising the new scoring system was 1.6% higher than the traditional system, with 88.6% agreement between the two approaches. Of particular interest were the borderline results, which were reviewed to ensure optimal performance of the new model in identifying cut-off scores between pass and fail outcomes. This led to further minor modifications to the new scoring grid. The final model resulted in a pass rate using the new scoring grid that was 2.27% lower than the traditional approach with an agreement of 91.8% between the two systems. The overall pass rate in 2018 using the traditional approach was 72.2%. In comparison, the new approach which was fully adopted in 2019 resulted in a pass rate of 72.4%. This difference of 0.2% was within the normal variability of results from year to year.

Review of the 2018 results also demonstrated an improvement in inter-rater reliability. Agreement on the same score between examiner pairs increased from 22.3% in the traditional approach to 73.6% in the new approach with a Kappa score of 0.7. In addition, examiner pairs reaching the same pass/fail outcome improved from 64.9% to 83.2% and the variability in scores between examiners reduced from an average of 1.6-point difference to a difference of 0.48.

## Communication and Training

The new changes to assessment and their impact were broadly communicated to all stakeholders including candidates, examiners and candidate supervisors. A short video was prepared by the College Censor outlining the new approach. Background documents, training materials and calibration materials were updated. The new approach utilising the

refreshed marking rubric and scoring grid was fully implemented and adopted in 2019. A final survey of both examiners and candidates following the 2019 examinations was again supportive of the refreshed assessment approach.

## Discussion

The RACP carries the education responsibility for physician training and assessment in Australia and Aotearoa New Zealand. With this comes the responsibility of the general community that trainees have met a consistently high standard of training and have the appropriate knowledge and skills of a physician. While trainees across the two countries train in diverse settings with a large number of supervising physicians, the clinical examination provides a consistent approach to summative assessment of clinical skills. Successful completion of examinations marks the end of basic physician training, which is then followed by advanced training across 33 sub-specialty training programs.<sup>7</sup> Thus, ensuring high standard foundation skills at the end of basic training is of high importance. Acknowledging that assessment drives learning, it is recognised that the clinical examination drives the attainment of these skills for all trainees across all training sites.

The clinical examination is a high stakes examination and currently there is little evidence in the literature guiding the selection of examination approach.<sup>8</sup> In this context, the current approach was not changed and the format of two long cases and four short cases was continued. The feasibility of the new marking and scoring systems was confirmed in the formal evaluation of the new approaches. The RACP is in line with national and international medical education bodies and is progressively adopting PA. However, given the large trainee numbers across multiple sites in two countries, it will take time and effort to embed fully and be reliable as an assessment approach. It may be argued that given the size of the RACP training program and the complexity of implementation of PA, that a high stakes point-in-time examination will appropriately continue to have a role in the assessment approach long term and will be an additional tool to guide learning through provision of meaningful feedback.<sup>9</sup> The review of the examination with the revised marking system has enabled improvements in the ability to provide meaningful feedback to trainees and may continue to have a place in the context of PA.

A review of scores confirmed long held concerns about the triple weighting of the long case and the simplistic approach to achieving a pass mark in the summation of points achieved across all cases. The evolution of the new compensatory system, which recognised the overlap in domains being assessed in short and long cases but also recognised their independence, was felt to be a superior approach and acceptability was rated highly by examiners and candidates. It reflected the holistic assessment process rather than a sum of individual parts.

The skills required of a physician are increasingly complex, contributed to by the ageing population and increasing rates of chronic disease. A physician needs to be able to holistically assess a patient and establish a comprehensive plan for investigation and management that is tailored to the individual patient taking into account outcomes, experience and costs.<sup>10</sup> The strength of long and short case examinations is that they not only involve assessment of accuracy of history taking and examination but also include assessment of complex skills integral to physician practice including the ability to synthesise, prioritise and formulate management plans for patients with complex medical and psychosocial issues. Reflecting the importance of these skills, the educational purpose of the clinical examination was strengthened by linking the marking rubric to the curriculum and clarification of the learning outcomes being assessed.

Recognising that a single assessment method will not assess all intended learning outcomes of a training program, the clinical examination cannot be considered in isolation from assessments which are already occurring as part of the implementation of PA but should be seen as a component in the totality of the assessment process. The validity of the clinical examination was improved by clarifying what aspects of physician practice are being assessed and linking the marking rubric including the domains for assessment to the examination purpose statements.

Reliability of long cases and short cases has previously been explored with findings that increasing examination time or supplementing with other summative assessments such as workplace assessments may increase the reliability.<sup>11</sup> Historically, reliability of the clinical examinations had been improved by doubling the number of cases and increasing the number of examiners involved in the assessments of individual candidates. It was not considered feasible to increase the number of cases any further as the examination already extends over an entire day for an individual candidate. Currently, examiner calibration is robust and annual calibration is mandatory to participate in the examinations. Using



paired examiners with scores being awarded through discussion and consensus results in less variability in scores.<sup>12</sup> Improved inter-rater reliability following simplification and standard clarification in the scoring rubric provided an additional increase in reliability of the examination.

Equivalence in examinations relates to whether or not trainees achieve comparable success regardless of the site at which they were examined. Upon review, the approach to this was considered robust. Examiners from the national panel travel intrastate and interstate to examine and a small number each year examine in both Australia and Aotearoa New Zealand. This has long been considered an important aspect of the exam to ensure equivalency between countries, states and individual sites.

The catalytic effect of an assessment refers to the impact of scores and feedback driving future learning.<sup>4</sup> It was recognised that the traditional 19-point scale impacted the ability of examiners to provide feedback as the nuances of scores were often difficult to interpret. Simplifying the marking rubric to a 6-point scale with clarification of expected standards for each score provided improved transparency and ability to provide robust feedback to candidates. The simplification of the scoring system also led to an increase in acceptability of the assessment process by both examiners and candidates.

## Conclusion

The refresh of the RACP clinical examination in 2015–2019 brought about many changes that improved the assessment approach with clarification of purpose, greater transparency of standards and an easier to use marking rubric. The new scoring system enabled a more holistic assessment of candidate performance. The role of the examination in the context of PA implementation will likely be reviewed in the future. We believe the improvements in the assessment approach as a result of this review fulfill the criteria for good assessment, including the Van der Vleuten principles and are applicable and generalisable to other clinical examinations.

## Ethics Approval

As this was a quality improvement activity related to RACP clinical examinations, no institutional review or ethics approval was sought. However, the manuscript was reviewed by the Executive General Manager, Education, Learning and Assessment, RACP and the Chair of the Education Committee, RACP who both provided input into the final manuscript.

## Acknowledgments

The authors wish to acknowledge the clinical examination review working group, the Royal Australasian College of Physicians Examination Unit and the National Examiner Panel for their contribution to the refresh of the clinical examination.

## Disclosure

The authors report no conflicts of interest in this work.

## References

1. Standards for RACP assessment programs. Royal Australasian College of Physicians website; 2016. Available from: <https://www.racp.edu.au>. Accessed January 23, 2024.
2. Govaerts M, Van der Vleuten C, Schut S. Implementation of programmatic assessment: challenges and lessons learned. *Educ Sci*. 2022;12(10):717. doi:10.3390/educi12100717
3. Schuwirth L, Van der Vleuten C, Durning SJ. What programmatic assessment in medical education can learn from healthcare. *Perspect Med Educ*. 2017;6(4):211–215. doi:10.1007/s40037-017-0345-1
4. Norcini J, Brownell A, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206–214. doi:10.3109/0142159X.2011.551559
5. Van der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39(3):309–317. doi:10.1111/j.1365-2929.2005.02094.x
6. Cizek GJ, editor. *Setting Performance Standards: Foundations, Methods and Innovations*. 1st ed. Routledge; 2001. doi:10.4324/9780203848203
7. What is a physician or paediatrician? Royal Australasian College of Physicians website. Available from: <https://www.racp.edu.au>. Accessed January 23, 2024.

8. Patterson S, Shaw L, Rank MM, Vaughan B. Assessments used for summative purposes during internal medicine specialist training: a rapid review. *Educ Sci.* **2023**;13(10):1057. doi:10.3390/educsci13101057
9. Touchie C. High-stakes point-in time assessment in the era of CBME: time's up? Ice Blog Admin; **2021**.
10. Combes JR, Arespachochaga E. Physician competencies for a 21st. Century health care system. *J Grad Med Educ.* **2012**;4(3):401–405. doi:10.4300/JGME-04-03-33
11. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. *Med Educ.* **2008**;42(9):887–893. doi:10.1111/j.1365-2923.2008.03129.x
12. Faherty A, Counihan T, Kropmans T, Finn Y. Inter-rater reliability in clinical assessments: do examiner pairings influence candidate ratings? *BMC Med Educ.* **2020**;20(1):147. doi:10.1186/s12909-020-02009-4

### Advances in Medical Education and Practice

Dovepress

### Publish your work in this journal

Advances in Medical Education and Practice is an international, peer-reviewed, open access journal that aims to present and publish research on Medical Education covering medical, dental, nursing and allied health care professional education. The journal covers undergraduate education, postgraduate training and continuing medical education including emerging trends and innovative models linking education, research, and health care services. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <http://www.dovepress.com/advances-in-medical-education-and-practice-journal>