Too hot to handle? Assessing the validity and reliability of the College of Intensive Care Medicine “Hot Case” examination

Kenneth R Hoffman, Christopher P Nickson, Anna T Ryan and Stuart Lane

The second part examination for the College of Intensive Care Medicine of Australia and New Zealand (CICM) is the second summative barrier examination of the 6-year training program. It is a quintessential high-stakes examination, requiring extensive study for 6–12 months and with a mean pass rate of 50–55%.¹ The examination comprises a written paper and then progression of successful candidates to an oral examination. The latter is composed of a viva examination and two clinical examination cases referred to colloquially as “Hot Cases”, performed on real patients in the intensive care unit (ICU). Each Hot Case lasts 20 minutes, consisting of an observed 10-minute clinical examination and a 10-minute case presentation and discussion with two examiners.

The Hot Case component has a mean pass rate of 64%, although only 37% of candidates pass both cases.¹ The Hot Case examination’s relatively low pass rate in comparison with the viva pass rate (75%)¹ highlights the potential disparity and the need for the ongoing evaluation of this component. This is particularly important given the potential impact of a false negative (candidates who fail and who should have passed) or a false positive examination result (candidates who pass and who should have failed) on both trainee wellbeing and community safety.

The CICM’s responsibility for credentialling

Credentialling trainees for specialist practice is a vital role of medical specialist colleges. The CICM Constitution states that the college will “determine and maintain professional standards for the practice of intensive care medicine in Australia and New Zealand”.² Health care education institutions combine teaching with credentialling and are accountable to the trainee, the patient and society.³ The focus of credentialling is protecting the patient and the community, necessitating clearly defined performance standards for trainees.⁴ High-stakes examinations, including the CICM second part examination, have an established history as the preferred mechanism to identify trainees who have met required standards.⁵ These examinations should be precise in specifying the requirement for competency

ABSTRACT

The College of Intensive Care Medicine of Australia and New Zealand is responsible for credentialling trainees for specialist practice in intensive care medicine for the safety of patients and the community. This involves defining trainees’ performance standards and testing trainees against those standards to ensure safe practice. The second part examination performed towards the end of the training program is a high-stakes assessment. The two clinical “Hot Cases” performed in the examination have a low pass rate, with most candidates failing at least one of the cases. There is increasing expectation for medical specialist training colleges to provide fair and transparent assessment processes to enable defensible decisions regarding trainee progression. Examinations are a surrogate marker of clinical performance with advantages, disadvantages and inevitable compromises. This article evaluates the Hot Case examination using Kane’s validity framework and van der Vleuten’s utility equation, and identifies issues with validity and reliability which could be managed through an ongoing improvement process.

Crit Care Resusc 2022; 24 (1): 87-92

and should verify the chosen performance standard is appropriate for safe practice.⁶

High-stakes examinations have important implications for candidates, requiring personal and professional sacrifices in preparation for the examination.⁷ Failure has major consequences, with emerging literature demonstrating failure can undermine confidence, affect wellbeing, disrupt professional and personal life, and may even demoralise some candidates into leaving the workforce.⁸

Health care education providers, including specialist colleges, are undergoing increasing expectation and requirement to define their acceptable standard for progression and provide fair and transparent assessment processes that allow defensible decisions regarding

trainee progression.⁹ The CICM is responsible for “support programs of training and education ... commensurate with specialist practice in intensive care medicine”.²

Evaluation of specialist college examination processes is required to ensure that the correct candidates progress and that false negative and false positive examination results are minimised. This requires an appropriate standard setting so the examination can dichotomise candidates into satisfactory and unsatisfactory. An example of ongoing improvement in the second part examination is the recent transition to the Angoff scoring method for the written paper. The Angoff method is a standard-setting process widely considered to be fair, validated, reliable and defensible.¹⁰ For high-stakes examinations, defining the passing standard requires a robust and ongoing assessment to validate decisions, as there is a continuous spectrum of candidate performance, with variable knowledge and clinical experience due to a broad syllabus being tested at a single time point.

Using educational theory to analyse assessments

Assessments are a surrogate marker of clinical performance used to test competence, which can be difficult to define. There is no perfect format, with every assessment having advantages, disadvantages and inevitable compromises.¹¹ Medical assessment beliefs often conform to institutional, cultural and local practice rather than an understanding of assessment theory.¹² Although education forms a critical component of a physician's role as a scholar,¹³ medical educators and assessors may lack formal training in education theory and research.^{3,14} Two internationally recognised frameworks, commonly used by health care education communities to evaluate validity and reliability in assessments, were used to evaluate the Hot Case assessment: Kane's validity framework¹⁵ and van der Vleuten's utility equation.¹² Kane's framework was selected because earlier validity frameworks have excessive types of validity or fail to prioritise sources of validity evidence. Kane's framework addresses both concerns by emphasising key inferences as the assessment progresses from a single observation to a final decision.¹⁶ Van der Vleuten's utility equation provides a conceptual model for assessments with practical implications, while avoiding complex psychometrics that often discourage health professional readers.¹¹

Validity argument for the Hot Case examination

Validity refers to measurement of clinical competence using an authentic measurement in the assessment process to enable defensible decisions regarding candidates' appropriateness for progression in a training program.¹¹

Validity refers to the interpretation of test results, rather than the test itself.¹⁵ Kane's validity framework outlines four components of validity: scoring (converting an observed performance into a score), generalisation (considering the score as a representation of performance in a test environment), extrapolation (using the score to represent real-world performance) and implications (applying the score for decision making).¹⁵ Making appropriate decisions about candidate performance requires an understanding of validity strengths and weaknesses, and evidence is required to verify an examination's use in discerning trainee performance. This article considers the Hot Case examination using all four components of Kane's validity framework.

Scoring

Scoring validity refers to the process of converting an observed performance into an examination score. It asks the question, “does the score accurately reflect the observed performance?” This involves aspects including question development, examiner standardisation and test security in addition to more nuanced aspects, including scoring criteria and expectations. Currently, trainees preparing for the Hot Case examination lack a clear scoring rubric outlining task expectations. The absence of scoring criteria impairs alignment of the observed clinical performance with the final assessment score for trainees. While examiners may reference internal scoring criteria, the absence of these explicit criteria for trainees can set unclear expectations. This can lead to information being derived from previous successful candidates, rather than clearly specified expectations from the assessment outline. Increasing transparency of the examination process by clearly outlining the areas being assessed and the associated scoring rubric used by examiners would enhance this aspect of validity. Examples would be the expected standard and criteria for assessment for the clinical examination, presentation and discussion. Scoring rubrics can also form a useful framework for providing candidates with clear feedback, whether they are successful or not.

Generalisation

Kane's generalisation validity asks, “does this examination test what we want to know?” Hot Case examination candidates are asked a clinical question after receiving a focused history. To answer the question, they must examine the patient, present their clinical findings and request further information to justify their overall answer. This aspect of Kane's validity framework is pragmatic, requiring candidates to demonstrate expertise in patient examination and clinical reasoning. Although the principles of what is being tested appear appropriate, Kane's

framework asks if the delivery of the assessment interferes with how these aspects are assessed. One consideration is whether the time allocated for the examination affects candidate performance, as perceived time constraints may lead to either overextrapolation of answers or excessive brevity. While examinations need to involve some form of time restriction, greater time allowances ensure that examiners can probe into areas where they are unsure of the candidates' answers, clarifying their understanding. Furthermore, it reassures candidates that they have time to answer a question fully, without feeling time-pressured and missing out vital information they may wish to discuss. The responsibility to progress the assessment at the required speed rests with the examiner not the candidate, and increasing assessment time allows for examiner and candidate performance to be optimised.

Extrapolation

Extrapolation validity asks, "does a good performance in the test reflect real world proficiency?" Given that Hot Case examinations are performed with real ICU patients, the extrapolation aspect of this framework appears strong, as examiners and candidates respect the task authenticity of integrating clinical skills with communication and professional competency.¹⁶ A potential concern here is that Hot Case examinations are a standardised measure of performance in a complex environment where the patients being examined have received optimal care from an interprofessional team, often for a significant time. Since aspects of validity are impaired by reductionist approaches to assessment, which divide complex tasks into objectively assessable smaller skills,¹² the Hot Case examination needs to ensure the expected performance of a candidate is what would be required of a real-world clinician at the same point in time. This aspect of validity appears strong, as the expected performance in the second part examination is a summative barrier for CICM trainees to progress into their final year of training.

Implications

Kane's final implications validity asks, "how does this decision impact the progression of the candidate in the assessment process?" The progression aspect of the Hot Case examination is significant and relevant, as the second part examination rules specify that the whole examination is failed automatically if the candidate fails both Hot Cases and receives a mark below 40% in the Hot Case component.¹⁷ These progression criteria apply regardless of candidates' performance in the written or viva sections, implying that proficiency in the clinical environment is seen as a vital aspect of being satisfactory in the overall examination.

This statement could be seen as face validity, whereby an assessment appears effective in terms of its stated aims. Kane's implications validity suggests that Hot Cases are seen as integral to being satisfactory in the overall examination. By necessity, Kane's scoring validity, generalisation validity and extrapolation validity must be optimised to ensure that the implications validity is achieved.

Kane's framework demonstrates that ensuring validity of an assessment requires attention to all parts of the process, not just the content itself. In practice, this means that for an organisation to maintain the validity of its assessments, it must continually review and amend the process as knowledge and understanding of assessment evolve. The CICM recognises this and recently requested and underwent a review of its assessment processes by the Australian Council for Educational Research as part of its ongoing requirement for accreditation as a specialist medical college by the Australian Medical Council.

Reliability argument for the Hot Case examination

Van der Vleuten's utility equation¹² views assessment through the variables of validity, reliability, educational impact, acceptability and cost. There is overlap between the utility equation and Kane's framework, particularly the closely related variables of validity and reliability.

Reliability measures how consistently an assessment measures performance and the impact of task, patient, examiner and assessment context.¹² Reliability primarily affects candidates closest to the pass/fail boundary, with those having clearly passed or failed being less affected by assessment consistency. Reliability is quantified using the reliability coefficient from 0 (completely unreliable) to 1 (completely reliable). For high-stakes assessment, the minimal acceptable value is 0.8, equating to a 20% false negative/false positive rate.¹¹ While increasing the structure of assessment with scoring keys and standardised protocols improves reliability, increasing the assessment sample size is the most effective method of improvement, as highlighted by the Royal Australasian College of Physicians. A single one-hour clinical long case reliability coefficient is 0.38, with 4–5 hours of assessment time required to achieve a reliability coefficient greater than 0.7.¹⁸ Decisions made based on assessment tasks should reflect the quality of information the task provides.¹⁹ In high-stakes assessment, increasing the assessment time maximises the final decision strength.²⁰

Theoretically, if an examination measures performance reliably, a candidate's individual Hot Case marks should correlate strongly. The reliability coefficient of the CICM Hot Case examination is not reported in the assessment outline and is likely to be low, based on the total assessment

time of 40 minutes. Generalisability theory is a statistical methodology that estimates an overall assessment of reliability plus analysis of the relative contributions of the candidate, case and examiner to the overall reliability.¹⁸ If the reliability coefficient is unacceptable after the statistical analysis is completed, increasing the total assessment time through more cases (currently two), more time per case (currently 20 minutes) or both can be considered. Hot Cases are performed with two examiners and an overseeing senior examiner. If examiners are well trained and candidates have different examiners per case, increasing the number of examiners would improve reliability less than increasing the overall assessment time.¹² In candidates with borderline performance, gathering more data by increasing the assessment time is required to justify the final decision.⁹ This is important given the significant consequences of failing the examination.

Candidates are currently required to perform four satisfactory formative Hot Cases before applying for the second part examination.²¹ These cases are performed with a CICM supervisor of training or delegated CICM Fellow. Including these cases in the summative process would increase the reliability of the Hot Cases. However, this would require significant CICM faculty development and review of the validity and reliability, as highlighted in this article. The process of using more frequently collected information about trainees' competence, providing feedback to inform trainees of performance, and allowing summative decisions to be based on multiple sources of information is an assessment paradigm called programmatic assessment.^{22,23} Programmatic assessment is gaining favour in health care education settings such as medical schools and specialist colleges.

The educational impact, acceptability and cost of Hot Case examinations

Educational impact describes the effect of assessment on learning. Trainees direct learning towards assessed material to optimise chances of academic success.²⁴ As stated by Sir William Osler:²⁵

I do not know of any stimulus so healthy as knowledge on the part of the student that he will receive an examination at the end of his course. It gives sharpness to his dissecting knife, heat to his Bunsen burner, a well worn appearance to his stethoscope, and a particular neatness to his bandaging.

To maximise the educational impact of a task, assessment should promote adaptive learning behaviour to prepare trainees for their professional role. Constructive use of assessment tasks underlies the transition from assessment of learning to assessment for learning,²⁶ and perhaps assessment as learning.

The educational impact of Hot Cases is high. Transition from novice to expert requires accumulation of knowledge reinforced by experience and the ability to apply problem-solving to clinical situations.¹² Maintaining Hot Case authenticity directs learning to competent clinical practice. Historically, formative workplace-based feedback during training was limited, although it is now incorporated. Hot Case practice provides an opportunity for direct observation of clinical examination skills, oral presentation skills, clinical reasoning through discussion, and time to reflect on performance from immediate feedback, rarely achievable in a busy ICU. Hot Case practice can be organised when trainees and examiners are not rostered clinically and candidates frequently practise Hot Cases in different hospitals, with varying perspectives from different examiners and experiencing less familiar ICU subspecialties. While this is motivated by passing a high-stakes examination, it also drives improvements in the candidate's current and future clinical practice.

Acceptability is the degree to which the relevant community trusts the assessment process¹² and is linked with implications validity. Trainees and examiners often do not question the examination process, unaware of the literature supporting or contradicting these practices.²⁷ Currently, within Australia and New Zealand, there is no best practice examination format established between specialist colleges, in contrast to medical schools via Medical Deans Australia and New Zealand. However, as assessment remains a core function of the CICM, applying evidence-based examination principles is a priority for maintaining academic standards and improving the training experience.

The final variable in the utility equation is cost because well designed assessments are resource-intensive for examiner training, data processing, monitoring and review. Trainees are more accepting of teaching costs than those of assessment.¹² Specialist examinations are expensive for trainees, although the CICM second part fee (\$3910 in 2021) compares favourably with other colleges. A screening written paper limits the progression to more resource intensive clinical components. The coronavirus disease 2019 (COVID-19) pandemic response has necessitated changes to the second part examination format, and while this did not reduce costs, the ability to deliver the assessment both online and locally was demonstrated. However, this may have ethical implications for examiners assessing their own local trainees, potentially compromising the examination validity and reliability. While there are advantages to using an online virtual format for examinations,²⁸ difficulties in assessing clinical skills may be considered unacceptable by candidates and examiners.²⁹

Conclusion

The CICM second part examination was the first examination of its kind in the world and has had a lasting impact internationally.⁵ Viewed through Kane's validity framework and van der Vleuten's utility equation, the Hot Case examination has significant strengths and obvious weaknesses. All assessments require ongoing re-evaluation over time, as clinical practice and education knowledge evolves. High-stakes assessment should be delivered with a concurrent quality assurance and continuous improvement process. This includes balancing the objective evaluation of the assessment process through the lens of education theory with the practical and pragmatic delivery of the assessment. The CICM is required to do this as part of its accreditation as a medical specialty college and has shown recent commitment to this in a review of its assessment processes by the Australian Council for Educational Research.

Specific concerns with the current Hot Case format should be addressed openly due to the high-stakes nature of the examination. The validity could be considered, aligning with Kane's framework to ensure scoring, generalisation, extrapolation and implication validity of the assessment. The reliability could be formally quantified with the reliability coefficient and if low, appropriate amendments to the examination process should be made to increase the assessment time. This would strengthen the areas of the examination that may be of concern and would lead to a clearer understanding of how the second part examination, and specifically the Hot Cases, could be embedded within a future CICM curriculum based on programmatic assessment.

Graduating CICM Fellows should provide the community with a high standard of care, and to ensure this, the CICM should provide trainees with transparent, evidence-based examination processes that produce valid and reliable results, within a culture of continuous improvement that can make defensible decisions regarding trainee progression. The CICM Hot Case examinations have currently stood the test of time and, with further analysis and evaluation, they can justify their place in the future training curriculum.

Competing interests

Stuart Lane is the current Chair of the Second Part Examination Committee for the College of Intensive Care Medicine of Australia and New Zealand (CICM). Christopher Nickson is a Member of the First Part Examination Committee for CICM.

Author details

Kenneth R Hoffman^{1,2}

Christopher P Nickson^{1,2,3}

Anna T Ryan⁴

Stuart Lane^{5,6,7}

- 1 Alfred Hospital, Intensive Care Unit, Melbourne, VIC, Australia.
- 2 Department of Epidemiology and Preventative Medicine, School of Public Health, Monash University, Melbourne, VIC, Australia.
- 3 Alfred Health Centre for Health Innovation, Melbourne, VIC, Australia.
- 4 Department of Medical Education, Melbourne Medical School, Faculty of Medicine Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia
- 5 Nepean Hospital, Intensive Care Unit, Sydney, NSW, Australia.
- 6 Sydney Medical School, University of Sydney, Sydney, NSW, Australia.
- 7 College of Intensive Care Medicine of Australia and New Zealand, Melbourne, VIC, Australia.

Correspondence: k.hoffman@alfred.org.au

doi: <https://doi.org/10.51893/2022.1.R>

References

- 1 College of Intensive Care Medicine of Australia and New Zealand. Second part examination: exam report; March-May 2019. Melbourne: CICM, 2019. https://www.cicm.org.au/CICM_Media/CICMSite/CICM-Website/Trainees/Assessments%20and%20Examinations/Second%20Part%20Exam/TE-SP1-MARMAY2019-Exam-Report-Final.pdf (viewed Aug 2021).
- 2 College of Intensive Care Medicine of Australia and New Zealand. Constitution [Internet]. CICM; 2015 [updated 2015 May; cited 2021 Aug 17]. https://www.cicm.org.au/CICM_Media/CICMSite/Files/Regulations%20and%20Constitution/C-Gen_Constitution.pdf (viewed May 2021).
- 3 da Silva Campos Costa NM. Pedagogical training of medicine professors. *Rev Lat Am Enfermagem* 2010; 18: 102-8.
- 4 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Lanham, MD: AERA; 2014.
- 5 Lee RP, Venkatesh B, Morley P. Evidence-based evolution of the high stakes postgraduate intensive care examination in Australia and New Zealand. *Anaesth Intensive Care* 2009; 37: 525-31.
- 6 Plake B, Wise L. What is the role and importance of the revised AERA, APA, NCME Standards for educational and psychological testing? *Educ Meas* 2014; 33: 4-12.
- 7 Hutchinson L, Aitken P, Haynes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Med Educ* 2002; 36: 31-55.
- 8 Klasen JM, Lingard LA. Allowing failure for educational purposes in postgraduate clinical training: A narrative review. *Med Teach* 2019; 41: 1263-9.
- 9 Lane AS, Roberts C, Khanna P. Do we know who the person with the borderline score is, in standard-setting and decision-making. *Health Prof Edu* 2020; 6: 617-25.

REVIEW

- 10 Karcher C. The Angoff method in the written exam of the College of Intensive Care Medicine of Australia and New Zealand: setting a new standard. *Crit Care Resusc* 2019; 21: 6-8.
- 11 van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005; 39: 309-17.
- 12 van der Vleuten CP. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996; 1: 41-67.
- 13 Hautz SC, Hautz WE, Feufel MA, Spies CD. What makes a doctor a scholar: a systematic review and content analysis of outcome frameworks. *BMC Med Educ* 2016; 16: 119.
- 14 Nelson MS, Clayton BL, Moreno R. How medical school faculty regard educational research and make pedagogical decisions. *Acad Med* 1990; 65: 122-6.
- 15 Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013; 50: 1-73.
- 16 Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. 2015. *Med Educ* 2015; 49: 560-75.
- 17 College of Intensive Care Medicine of Australia and New Zealand. Notes to candidates for the second part examination. Melbourne: CICM, 2019. https://www.cicm.org.au/CICM_Media/CICMSite/Files/Training/T-18-Notes-to-candidates-for-the-Second-Part-Examination-2019.pdf (viewed Aug 2021).
- 18 Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. *Med Educ* 2008; 42: 887-93.
- 19 Dijkstra J, Galbraith R, Hodges BD, et al. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Med Educ* 2012; 1712: 20.
- 20 van der Vleuten CP. Revisiting "Assessing professional competence: from methods to programmes". *Med Educ* 2016; 50: 885-8.
- 21 College of Intensive Care Medicine of Australia and New Zealand. Hot (ICU) case assessment form. Melbourne: CICM, 2019. https://www.cicm.org.au/CICM_Media/CICMSite/Files/Exams/Hot-Case-Assessment-Form-ICU.pdf (viewed Aug 2021).
- 22 Schuwirth L, van der Vleuten C, Durning SJ. What programmatic assessment in medical education can learn from healthcare. *Perspect Med Educ* 2017; 6: 211-5.
- 23 Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012; 46: 38-48.
- 24 Shepard LA. Psychometricians' beliefs about learning. *Educ Res* 1991; 20: 2-16.
- 25 Silverman M, Murray T, Bryan C; editors. The quotable Osler. Philadelphia: American College of Physicians, 2002.
- 26 Black P, McCormick R, James M, Pedder D. Learning How to Learn and Assessment for Learning: a theoretical inquiry. *Res Pap Educ* 2006; 21: 119-32.
- 27 Nelson M, Clayton B, Moreno R. How medical school faculty regard educational research and make pedagogical decisions. *Acad Med* 1990; 65: 122-6.
- 28 Hess BJ, Kvern B. Using Kane's framework to build a validity argument supporting (or not) virtual OSCEs. *Med Teach* 2021; 43: 999-1004.
- 29 Hannon P, Lappe K, Griffin C, et al. An objective structured clinical examination: from examination room to Zoom breakout room. *Med Educ* 2020; 54: 861.