Health Professions Education

Volume 6 | Issue 4 Article 14

2020-12-01

Do We Know Who the Person With the Borderline Score is, in Standard-Setting and Decision-Making

Andrew S. Lane
Sydney Medical School, University of Sydney, Australia

Christopher Roberts
Sydney Medical School, University of Sydney, Australia

Priya Khanna
Sydney Medical School, University of Sydney, Australia

Recommended Citation
Lane, Andrew S.; Roberts, Christopher; and Khanna, Priya (2020) "Do We Know Who the Person With the Borderline Score is, in Standard-Setting and Decision-Making," *Health Professions Education*: Vol. 6: Iss. 4, Article 14.
DOI: 10.1016/j.hpe.2020.07.001
Available at: https://hpe.researchcommons.org/journal/vol6/iss4/14

Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 6 (2020) 617–625

www.elsevier.com/locate/hpe

Do We Know Who the Person With the Borderline Score is, in Standard-Setting and Decision-Making

Andrew S. Lane [a,*], Christopher Roberts [b], Priya Khanna [c]

[a] *Coordinator of Clinical Studies Sydney Medical School, University of Sydney, Australia*
[b] *Head of Faculty Development Sydney Medical School, University of Sydney, Australia*
[c] *Senior Lecturer in Curriculum Development, Sydney Medical School, University of Sydney, Australia*

Received 8 July 2020; accepted 9 July 2020
Available online 29 July 2020

Abstract

Purpose: When assessing clinical competence, health professional educators use assessments of knowledge attainment, skills acquisition, and professional development, which impact on decision-making for student's training progression. Given the impact of progression-failure, it is critical that the expected standard of performance is derived accurately, fairly, and transparently, and that the rating of student performance is performed within the highest standards achievable. There is ongoing disagreement as to the most appropriate methods to address both standard setting and decision-making. The borderline candidate has been debated extensively in the academic and educational setting, with ongoing disagreement surrounding the concept.
Methods: In this paper, we discuss further perspectives on the use of the borderline candidate, as part of the process for standard-setting, to give insights into how we can reframe the concept more accurately and apply it more appropriately.
Discussion: Drawing parallels to Kane's validity framework, we consider the concept of the borderline candidate from four different perspectives: 'what is'-what are the linguistics and implications behind the phrase 'borderline candidate'; 'who is'-who is the borderline candidate; decided 'by whom'-who is the person making the judgement; and 'under what circumstances'-the context of the assessment.
Conclusion: Finally, we translate the theoretical discussion into pragmatic and practical solutions in standard-setting practice

Keywords: Angoff method; Borderline candidate; Standard setting

1. Introduction

With the increasing impetus towards competency-based education, healthcare educational institutions and assessment bodies are under greater scrutiny than ever before to ensure defensibility, fairness and transparency in both defining the expected standard of performance, as well as rating student's performance for progression decisions. When setting the standards for

* Corresponding author. Nepean Clinical School, Derby Street, Penrith, NSW, 2750, Australia. Fax: +0247341811.
E-mail address: Stuart.lane@sydney.edu.au (A.S. Lane).
Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region.

https://doi.org/10.1016/j.hpe.2020.07.001

attainment of academic and professional competence, health professional educators use a wide variety of methods and tools to assess and document the learning progress, skill acquisition, readiness for progression, and dealing with specific educational needs of students.[1] Therefore, the academics and educators responsible for setting the standards need to have a clear understanding of nuances of methodological approaches and tools for assessing the attainment of these standards. They also need to avoid two key judgement errors – passing incompetent healthcare students or-trainees/professionals and failing competent healthcare students or-trainees/professionals. Standard-setting is an integral aspect of any assessment system that involves a range of stakeholders including policy makers, test developers, and measurement specialists to ensure that the test results will be meaningful and defensible.[2] The extant literature, however, suggests there is little agreement as to the most appropriate methodological approach to standard setting and decision-making to address both aspects fairly.

Methods for setting standards can be described broadly as either test-based or examinee-based. In test-based methods such as the Angoff[3] and Ebel[4] methods, judges review test items or prompts and estimate the expected level of performance of a borderline examinee (one just at the margin between two categories) on a given task. The patient-safety method similarly reviews performance test items (e.g., checklist items), to determine those that must be performed correctly to accomplish patient safety or other critical goals.[5] In examinee-based methods represented by the borderline group, judges categorize the performance of individual examinees, either through direct observation, review of proxies of their behaviour such as performance checklists, or review of examinee products such as chart notes written after a standardized patient encounter.[6] In these methods, the scores of examinees in different performance categories are utilized to generate the final cut score. Finally, compromise methods such as the Hofstee method combine features of absolute and relative standards, asking judges to estimate both acceptable passing scores and acceptable fail rates.[7]

Currently there is no uniformity in good practice in standard setting methods.[8] Most standard-setting methods pivot on the idea of the borderline or minimally competent student or examinee. A critical component of standard setting process is the setting of a 'cut score', defined as the point on a scoring-scale that separates one performance standard from another. There is no method which provides a 'gold standard' in determining a cut-score value for any assessment. Institutions are advised to select and implement a rigorous process by which a cut-score value can be arrived at, with appropriate supporting documentation and empirical evidence, that is defensible to stakeholders.[9] The most common divisions of assessee scores include satisfactory/unsatisfactory or pass/fail. In this situation, the cut score separates those who know, or (can show or do) just enough to pass from those who do not know (or show or do) enough to pass. Claims of exactness in dichotomising attainment of complex skills may lead to potentially erroneous conclusions about the grading or ranking of candidates undergoing an assessment, and subsequent progression decisions based on these assessments. Errors in making high stake progression decisions can have significant sequelae when considered from the viewpoint of patient safety and outcomes. The 'entity' that is most impacted with the dichotomous decision-making process is the 'borderline candidate.' That is, the candidate whose performance is exactly on the border between two performance categories.

Operationalising this definition of the borderline candidate in practice can be a difficult concept for many assessors and students to understand and apply in practice.[10] Educators would benefit from an explicit definition of what the concept of the borderline candidate is, and how it should be applied to their assessments in their context. The literature summarises some of the challenges. Friedman Ben-David states that ideally candidates should demonstrate mastery of competence by responding correctly to the task criteria and by achieving the maximum score. However, candidates' performances may vary according to their aptitude for differing tasks, an issue of context specificity. A requirement of a mastery approach to performance (fully competent) for passing each task may appear unrealistic in most situations.[10] A second issue in identifying borderline candidates is that many methods require the application of judgement, raising the issue of assessor subjectivity.[11] Both context-specificity and assessor-subjectivity undermine certainty for educators and assessors in accepting the concept of the borderline candidate. Whilst on one hand, there are calls for acknowledging the inherent subjectivity in decision-making process especially considering emerging assessment practices such as programmatic assessment, t, may create dissonance with the way 'borderline candidates' have been viewed traditionally within the assessment systems using pass/fail dichotomous outcomes.

In this paper, we aim to conceptualise the concept of the borderline candidate more accurately for all

stakeholders with a vested interest in assessment and suggest ways of applying it more appropriately. We will consider the borderline candidate from four different perspectives:

- 'what' – determines the use of language and meanings behind the phrase 'borderline candidate'.
- 'who' - who is the borderline candidate.
- 'whom' is it decided by – who is the person making the judgement.
- 'where' and 'under what circumstances' is the judgement being made in the context of the assessment.

In considering these four aspects of the borderline candidate, we draw parallels with Kane's framework of validity in assessment as given in (Fig. 1). Finally, in this article we aim to translate the theoretical discussion into pragmatic and practical solutions for all stakeholders in standard setting practice.

In this paper we use the word assessee to describe the person undergoing the assessment, which can refer to a student or a healthcare profession trainee, and we use the word assessor to describe the person assessing the assessee, which can refer to an examiner or educator. In this discussion paper we make the assumption that the validity of the assessment is not being questioned, whilst also recognising that validity itself is a term which can have many versions and interpretations,[12] and we will focus more on explicitness of terms, since explicitness aids in but does not provide validity on its own.[13]

2. Discussion

2.1. What do we mean by the actual words 'borderline candidate'?

In considering the meaning of, 'borderline' and 'candidate', the question arises as whether these words allow an accurate visualisation for assessors as to what they are required to do in standard setting exercises. Firstly, the word borderline. Assigning an assessee to a category of 'borderline' in assessment usually implies that the assessors are unsure if the assessee is neither clearly satisfactory nor clearly unsatisfactory.[14] A usual outcome from this scenario is that the assessors require the person to undergo further assessment while ensuring that the passing of incompetent students/trainees and failing competent student trainees is kept to the minimum within resources available.[15]

This can lead to confusion for assessment committees, as there is a difference between the process of using the 'borderline candidate' method in a standard-setting procedure for future decision-making, versus somebody who has achieved a 'borderline' (close to the cut-score' score in an assessment).[14] The former is a proactive decision, and the latter is a reactive decision. Let us take an example to see the interplay of these differences in a typical standard setting exercise. Consider a written short answer examination paper of fifteen questions, each question worth ten marks each, using the Angoff method for standard setting, the most commonly used method for written assessments within the UK[16] The Angoff method requires the assembly of a group of subject matter experts (SMEs), and using the

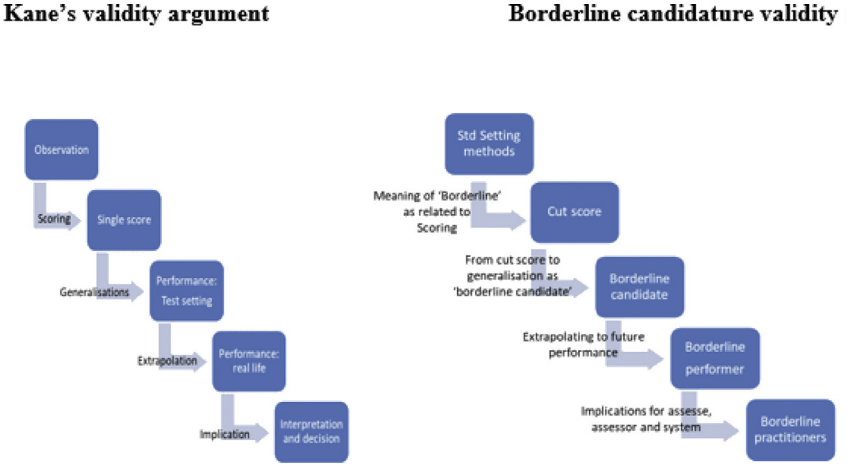

Fig. 1. Parallels between Kane's validity inferential framework and borderline candidature.

borderline candidate principle, are asked to evaluate each question and estimate the proportion of minimally competent examinees that would correctly answer the item. The ratings are averaged across raters for each item and then summed to obtain a 'cut-score' for the whole exam paper.[17] If a candidate were to achieve a score close to the 'cut-score' for the examination, it does not make them a 'borderline candidate' in the same context as the 'borderline candidate' that was used in standard-setting to determine the 'cut-score' for this examination. It means that they achieved a score close to the 'cut-score', and therefore assessors might consider their score a 'borderline' exam performance. In the same vein, a candidate were to score the calculated Angoff mark on one of the fifteen questions, once again it does not mean they are a borderline candidate in the same context as the 'borderline candidate' that was used to set the 'cut-score', it means they performed to the level of the theoretical borderline candidate on that question alone, since their overall exam performance may be well above or well below the 'cut-score' for the whole exam paper. They may achieve the cut-sore itself, however, this would mean they have achieved a satisfactory performance in the assessment, and although people might state their performance was borderline, it neither means they will be reassessed nor that they are a borderline candidate. Furthermore, from the perspective of a medical student or junior doctor assessment, it does not imply they are either a currently or potentially clinical practitioner whose practice might be seen by many colleagues as borderline; it simply applies to the assessment they have just completed.

Since the borderline candidate is often difficult to conceptualise when used in standard-setting, there have been recent attempts to make it easier by changing the nomenclature, however these attempts also have problems. To avoid the potential difficulties of the borderline candidate definitions,[18] alternative phrasing has been offered in the literature such as "just qualified candidate." [19] This phrase, however, can be contradictory, since if the candidates have just qualified, it implies that they have been satisfactory in all their assessments and may be about to commence practice. The principle of standard-setting applies to specific single assessments, and not overall qualification. This visually gives an examiner an impression of somebody who is already practicing, not somebody who is being assessed for practice. The same is true, for the term 'minimally competent candidate',[20] that also has implications to either to current or future work-performance rather than the assessment itself. Although this does not imply satisfactory completion of a range of assessments like the term 'just qualified,' it still gives an image of performance in a specific task?

These two alternative descriptions may be more relevant if the assessment is task-focused, from which future performance can be reasonably predicted. However, this term may not be suitable for decision-making around complex competencies that are situated and self-regulated. Therefore, the performance that is being considered as 'borderline' within standard setting is related to the assesse's performance within the examination. Hence, we propose reconceptualising borderline candidate in the context of standard-setting as 'borderline examination performer'.

However, for the purposes of assessments that are not task-focussed, this wording confuses the task of standard setting. What we are really looking at is performance in the specific assessment, therefore, if this is what we actual considering when we are setting the standards, then we should explicitly state it – the standard-setting is around examination performance, so what we are needing to conceptualise is a 'borderline examination performer'.

2.2. *Who is the 'borderline candidate'? - The persons and the image*

Another difficulty that assessors have difficulty when they try to use the concept of borderline candidate in standard-setting in viewing the borderline candidate as a person. Whereas the term refers to a concept, or more accurately a construct. A construct is different from a concept, in that a concept is an abstraction from some phenomenon that can be observed, while a construct is not something that you can observe.[21] In essence, a construct is a theoretical concept, making the idea of the borderline candidate even more difficult to work with and visualise in one's imagination.

Referring to the Angoff method as an example of a standard-setting using a borderline candidate method, assessors are meant to visualise a group of borderline candidates, and ask themselves the question 'what percentage of these borderline candidates would get this question correct or be satisfactory on this question?[22] Therefore, we are asking the assessor to decide on a satisfactory percentage of the group and not a single entity, which implies the imagined construct must have multiple entities. Therefore, within the group the borderline candidates will have varying performances based on their knowledge and application. However, if the assessor imagines a single borderline candidate to get an idea of who they are, and then imagines a who group of the same borderline

candidate they have just imagined, this does not lead to variability amongst the group of borderline candidates, simply multiple versions of the single borderline candidate. Based on the imagined qualities and performance of this single borderline candidate, this may contribute to either an over or underestimation of the borderline candidate's standard.

If assessors imagine one borderline candidate first, this also carries the risk of aligning the borderline candidate to specific examples of people they have known or currently know, and this is usually person they know who neither impresses them nor distresses them, in a sense they are average performers. There are two difficulties here, the concept of average, and the concept of the performer. When people consider the average performer, this is a norm-referenced concept.[11] However, the observed average assessee can vary greatly between different environments, tasks, and contexts, and it is performance of the tasks in that environment or context that are being observed, not the setting of a standard for a specific exam. When considering the borderline candidate, this should be a criterion-referenced construct - we cannot observe this,[11] and assessors must have a clear shared-idea of what they are, irrespective of our local working or academic environments.

By visualising a group of borderline candidates, of similar characteristics, there will be varied levels of knowledge and subsequent performance within the group, and it follows that this can vary on different days and in different environments. The borderline candidate is not a candidate in difficulty, since on some days they would be more likely to get the question correct than on others. Using the Angoff method again as an example, the effect of framing the borderline candidate as unable to perform well on a question *per se* impacts the overall cut-score of the assessment task. If the borderline candidate is seen as a poor performer, this will lead to lower marks being stated by the assessors for questions, meaning an overall reduction in the cut-score, and therefore a decrease in the standard required to pass, and ultimately more candidates passing the assessment than probably should.

With regard to the borderline candidate being seen to have a variety of performances on various days, there has to be a further acceptance that they are more likely to score well on topics that are easier to comprehend, and also repeatedly taught and asked. This means that certain topics which are considered by curriculum developers and assessors to be more integral to the core curriculum, are often taught with subjectively greater importance and practically with greater resources of time, and personnel, are more likely to be conducive to a satisfactory performance by a group of borderline candidates.

One way to potentially navigate this difficulty may be through greater use of the Ebel method for standard-setting, for written assessments? which asks two further questions beyond the Angoff of method when using the borderline candidate for standard-setting. With the Ebel method, there is a requirement to categorise the question as high, medium, or low difficulty, along with a scale of importance such as essential, important, and desirable.[19] This forces the assessor to consider the context of the question being asked, and therefore will consider the importance as well as the difficulty of question with reference to the overall syllabus.

2.3. The borderline candidate as decided by whom? – the impact on the assessor

There are specific nuances that the assessors and assessment boards must take into consideration when they are setting standards using the borderline candidate construct such as inherent biases on the part of assessors that need to be acknowledged in order to be minimised. Whilst the assessors are experts in their field, this does not make them free from bias, and rather than attempt to make the process bias-free, we should accept that they have biases and make sure their biases are minimised.[23] The whole point of having expert assessors is so that they can assist with setting the standard based on their expertise in the area. Specific biases that may come into play with assessors include areas of personal interest and personal belief regarding the importance of a topic. This is minimised by having a panel of experts who are likely to have a variety of interests, and clearly framing the expectations at the start of process as to the expected level of performance of the persons being tested, and that it is appropriate for the context of the assessment.

Whilst the implications of an assessment are mostly for the person being assessed, there are also implications also for the assessors conducting the assessment, and the assessors who are making a subsequent decision based on the result of the assessment. If the assessee is deemed to not be satisfactory in the assessment, it will likely have a negative consequence on their progress to some degree. However, it may be beneficial in the long-term, since if the assessors did not make the required non-satisfactory decision at the

time, then the candidates may well be put into a situation which they are not qualified to manage, and that situation obviously could affect the lives of others in a negative manner. Examples here would be a final-year medical student assessment, or an exit-qualification exam for a medical practitioner that grants them independent practice.

The effect on the assessor, although it may be to a lesser extent, can also be positive or negative, depending on whether they view the outcome of the assessment from the specific perspective of the assessee, or from the global perspective of the outcome of the assessment. We already know that many assessors feel uncomfortable when they award a 'fail grade' compared with a 'pass grade',[24] therefore assessors need to be clear what the consequences of their decisions are, and be clear on the responsibility associated with their decision. The 'failure to fail' is a recognised phenomenon, and it causes conflict for assessors, since by giving a negative grade to a student, the educator admits to having failed to effectively teach, motivate or create a learning environment for a particular student; however by unjustly giving a positive grade to a student the teacher does not ensure the quality of future patient care.[25] However, aside from the personal conflict that allows this situation to occur, the assessor may not have been given explicit instructions on the pass/fail criteria, and the issue is one of appropriate faculty development, as well uncertainty about the remediation process and its outcomes.[26] Assessors need to be clear if their duty to the assesse, or to the community, society, and their profession?[26] This is the human element: we believe that we need assessments that are robust enough tell us what we want to know, for the global benefit of what we are trying to achieve, but we also want them to be fair, for the specific benefit of the assessee. If there are inconsistencies between the assessors, in how the examination is conducted and how the assessment criteria are applied, then this affects reliability, which decreases the rigour of the assessment no matter how valid it might have been seen to be.

2.4. Under what circumstances? – the context of the assessment

The stakes of the examination should be taken into consideration when setting standards, and this should be based on the level of concern that there would be, should the assessment be incorrect. In any quality assurance exercise, there needs to be a minimisation of persons passing who should not have passed (false positive), and persons not passing who should have passed (false negative).[27] However the bigger picture needs to be taken into consideration. For high stakes progression decisions at the end of the training, assessments are more important than the ones during the middle of the training, since during the middle there are still opportunities for further assessment and remediation if required. Once the assessee has graduated this may not be possible. Hence, when there is a trade-off between false negatives and false positives because of statistical uncertainty, then the level of certainty that the person should have passed has to be higher. This does not mean that the actual process for setting the standards of the actual questions is altered, but the overall level of statistical certainty is increased. Therefore, if an exam had a cut score of 56%, with a standard error of the mean (SEM) of ± 1%, then it may be reasonable to have a cut score of 55% when the stakes are lower (lower end of the SEM range), and 57% when the stakes are higher (high end of the SEM range). The higher the stakes, the more robust the process is required to be to ensure both fairness and robustness of the process of setting and assessing standards.

2.5. Linking Kane's validity framework to standard setting

Our reconceptualization of issues related to borderline candidate and borderline performance in terms of what, who, whom and where seems analogous to Kane's validity framework.

Validation, in general, can be defined as the process of collecting and interpreting evidence to support the decision. Kane's framework emphasises key inferences as the assessment progresses from a single observation to a final decision. Kane's approach to validity can be applied to analyse any scenarios that involve articulating the claims and assumptions associated with the proposed decision.[28] Kane asserts validity is not a property of a test, but a property of the proposed interpretations and uses of test scores.[29] This reinforces that Kane is not looking at the validity of the assessment, but the development of assessment criteria. This frames the validity of an assessment not only when it is designed, but more importantly when it is delivered, bringing assessors into the influence on the eventual outcome. In Fig. 1, we have drawn parallels between the sequential framework put forward by Kane, with the four stages outline, and our own discussion of the concept of the borderline candidate.

The four types of inferences in Kane's framework, namely,

- Scoring: standard setting tools, processes, and procedures such as cut scores that define borderline performance (what is borderline).
- Generalisation: from a single cut score to a person as borderline.
- Extrapolation: who (judges) and why (circumstances) for inferring future performance.
- Implications: for the assessor, assessee, and system in terms of borderline practitioner.

Kane's framework provides a sound theoretical scaffolding for their everyday use in assessment and ask that any claims about a learner's performance are supported by appropriate evidence. Validation therefore consists of a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard. The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version of the desired level of competence. In much of the literature on standard-setting, the distinction between the passing score and the performance standard is not explicitly drawn, making it difficult to evaluate the validity of the interpretation assigned to the passing score. Maintaining a clear distinction between the passing score and the corresponding performance assists in this regard.

Moves towards introducing newer forms of assessment practice have compounded the uncertainty and difficulty surrounding standard-setting using the borderline candidate. Even within the prevailing assessment practices of having a summative, high-volume, high-stakes examinations, there is uncertainty around who has passed and who has failed. Therefore, with multiple, lesser-volume, low-to-medium stakes assessments, encompassing rich-narrative feedback, such as the method of programmatic assessment as suggested by Schurwith and Van De Vleuten, might allow the use of expert opinion for assessment for learning versus the assessment of learning.[30] Programmatic assessment is a summation of assessments rather than a summative assessment, with each assessment giving rich narrative feedback to the student on where they can improve rather than awarding a binary pass-fail decision. It is the rich narrative feedback that makes the decision for an assessee's progression requirement for learning clear, as the program of assessment gives multiple expert opinions from multiple assessments. This translates to a progression decision that is very defensible, but also provides constructive feedback for the assessee, promoting agency in learners such being lifelong reflective learners What we have tried to achieve in the article is help assessors better understand the construct of the borderline candidate, meaning they will be able to assess assesees as ready for purpose with greater confidence and fairness.

3. Recommendations and conclusion

In this paper, we have outlined the current difficulties with the use of the borderline candidate as used for standard-setting and attempted to reconceptualise the concept of and implications of borderline candidate from multiple perspectives. To assist assessors in utilising the borderline candidate principle optimally we suggest five steps should be taken.

1. Nomenclature: The term borderline candidate for standard-setting should be replaced by the term borderline examination performer. This language ensures that the assessor considers the context of the specific examination that they are setting the standards for, and that it is about performance in the examination, and not decision-making around practice before or after the examination.
2. Construct re-visualisation: When assessors are considering the borderline examination performance to set standards, they should consider a group of candidates and not a just one. This will avoid the assessor imaging one borderline examination performer and multiplying them, which may be inaccurate.
3. Construct ability: The borderline examination performance can vary on different days, and this performance should be better associated with those topics/competency areas that have had greater exposure and preparation in the curriculum.
4. Revisiting examiner standardisation: Assessors are the experts in their field, however there their biases and beliefs should be taken into consideration in an collegiate,open, and transparent manner when establishing the expectation of performance, so as to minimise bias as much as possible.
5. Re-contextualised statistical variability: The context of the stakes of the examination should

be considered, and although it does not affect the standard-setting process, it may adjust the statistical certainty required for an overall cut-score.

Considering emerging views on assessments of complex healthcare competencies and validity framework, the concept of borderline candidate needs to be revisited including standard settings methodologies. By reconceptualising the tensions around the concept of borderline candidate from perspectives of assessor, assesse and structural frameworks (such as standard setting), valid and robust inferences around future performance of such candidates can be undertaken.

4. Final considerations for readers

- Standard setting is an imprecise art yet significantly impacts decisions about candidates' progression to the next stage of training.
- The term borderline candidate is imprecisely defined in the literature and in common practice.
- We have examined the construct of the borderline candidate from four perspectives: what, who, whom and under what circumstances.
- Examiner briefing about standard setting should include: defining the meaning of the terminology, reframing the construct including the expected examinee ability, ensuring assessors reflect on their potential biases, and erring towards the candidate or to protecting the public depending on the stakes of the assessment.

Ethical approval

Ethical approval was not required for this manuscript.

Funding

No funding was received or required for this manuscript.

Declaration of competing interest

The authors have neither financial nor academic conflicts of interests.

References

1. The glossary of education reform. Retrieved 4th May 2019 https://www.edglossary.org/assessment/.
2. Norcini J, Guille R. Combining tests and setting standards. *Int Handbook Med Res Educ*. 2002;7:811–834.
3. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. *Educational Measurement*. Washington, DC: American Council on Education; 1971:508–600.
4. Ebel R. *Essentials of Educational Measurement*. Prentice-Hall; 1972.
5. Yudkowsky R, Park YS, Riddle J, Palladino C, Bordage G. Clinically discriminating checklists versus thoroughness checklists: improving the validity of performance test scores. *Acad Med*. 2014;89:1057–1062.
6. Livingston SA, Zieky MJ. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. 1982.
7. Hofstee WKB. The case for compromise in educational selection and grading. In: Anderson SB, Helminck JS, eds. *On Educational Testing*. San Francisco: Jossey-Bass; 1983:109–127.
8. Kampe N, Wagner H, Koller O. The standard-setting process: validating interpretations of stakeholders. *Large Scale Assess Educ*. 2019;7(3):71–78.
9. De Champlain AF. *Standard Setting Methods in Medical Education: High-stakes Assessment. Understanding Medical Education. Evidence, Theory, and Practice*. 2nd ed. John Wiley & Sons, Ltd; 2014.
10. Friedman M. AMEE Guide No. 18: Standard setting in student assessment. *Med Teach*. 2000;22(2):120–130.
11. Hejri SM, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. *Med J Islam Repub Iran*. 2014:1–6.
12. Royal K. Four tenets of modern validity theory for medical education assessment and evaluation. *Adv Med Educ Pract*. 2017;8:567–570.
13. Brown G. *Assessing Student Learning in Higher Education*. Taylor and Francis Ltd; 1997.
14. Sturmberg JP, Hinchy J. Borderline competence–from a complexity perspective: conceptualization and implementation for certifying examinations. *J Eval Clin Pract*. 2010 Aug;16(4):867–872.
15. Cusimano MD. Standard setting in medical education. *Acad Med*. 1996 Oct 1;71(10):S112–S120.
16. George S, Haque MS, Oyebode F. Standard setting: comparison of two methods. *BMC Med Educ*. 2006;6, 46.
17. Tannenbaum RJ, Kannan P. Consistency of angoff-based standard-setting judgments: are item judgments and passing scores replicable across different panels of experts? *Educ Assess*. 2015 Jan 2;20(1):66–78.
18. Zieky MJ. So much has changed: how the setting of cut-scores has evolved since the 1980s. In: Cizek GJ, ed. *Setting Performance Standards*. Mahwah, NJ: Lawrence Erlbaum Associates; 2001:19–52.
19. Psychometrics and Assessment Services of the Medical Council of Canada. *Technical Report on the Standard Setting Exercise for the Medical Council of Canada Qualifying Examination Part II*; 2015. Ottawa (CA): [Internet] [cited 2020 Mar 1]. Available from: https://mcc.ca/media/MCCQE-Part-II_Standard-Setting-Report_July-2015.pdf.

20. Burr SA, Zahra D, Cookson J, Salih V, Gabe-Thomas E, Robinson IM, et al. Angoff anchor statements: setting a flawed gold standard? *AMEE Med Educ Publ*. September 2019.
21. Leggett A. *Constructs, Variables and Operationalization*. 2011. Hair Marketing research. Ch 3 – Thinking like a researcher.
22. Homer M, Darling J. Setting standards in knowledge assessments: comparing Ebel and Cohen via Rasch. *Med Teach*. 2016;38(12):1267–1277.
23. Kukuckaa J, Kassin SM, Zapf PA, Dror IE. Cognitive bas and blindness: a global survey of forensic science examiners. *J Appl Res Memory Cogn*. 2017;6(4):452–459.
24. Van der Vossen MM. 'Failure to fail': the teacher's dilemma revisited. *Med Educ*. February 2019;53(2):108–110.
25. Yepes-Rios M, Dudek N, Duboyce R, Curtis J, Allard RJ, Varpio L. The failure to fail underperforming trainees in health professions education: a BEME systematic review: BEME Guide No. 42. *Med Teach*. 2016;38(11):1092–1099.
26. Wilkinson TJ, Tweed MJ, Egan TG, Ali AN, McKenzie JM, Moore M, et al. Joining the dots: conditional pass and programmatic assessment enhances recognition of problems with professionalism and factors hampering student progress. *BMC Med Educ*. 2011;11:29.
27. Homer M, Pell G, Fuller R. Problematizing the concept of the "borderline" group in performance assessments. *Med Teach*. 2017;39(5):469–475.
28. Cook DA, Brydges R, Ginsburg S, Halata R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015 Jun;49(6):560–755.
29. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50:1–73.
30. Van der Vleuten C, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–214.

A/Prof **Andrew Stuart Lane** is Coordinator of clinical studies in the Sydney Medical Program, and a Senior Staff Specialist in Intensive Care Medicine at Nepean Hospital, Sydney, Australia.

A/Prof **Christopher Roberts** Head of Faculty Development in the Sydney Medical Program, and an academic General Practitioner at Northern Clinical School, Sydney, Australia.

Dr **Priya Khanna** is a Senior Lecturer in curriculum development at Sydney Medical School, Sydney, Australia.