

Scaling Standard Set Marks to University Grade Boundaries in Health Education

Samantha L Strong

Aston University
s.strong2@aston.ac.uk

Amy L Sheppard

Aston University

Abstract: Assessments in higher education healthcare programmes can be challenging because they not only need to be fair, valid, and transparent, but it is also necessary to gauge safety, practical skill competence, and professionalism. One way to help maximise validity in practical assessments is to utilise ‘standard setting’ which aims to set a fair ‘cut score’ (pass mark) that reflects the score expected of a ‘minimally competent’ candidate. This purports that if a hypothetically minimally competent candidate could achieve this score, then it should be equivalent to the minimum pass mark for that assessment i.e. everyone who performs better than that should pass. This is effective from an assessment design point of view, but then leads to the new challenge that the cut score is unlikely to be equivalent to the university-level pass mark of 40% (or 50% for postgraduate), which, in cases where the component carries ‘weighting’, can lead to differences in pass mark across different assessments within the same module, or the same programme, which requires a way of scaling the marks for each assessment to make the cut score align with the university-level pass mark. However, the guidance for this is limited and unclear, so we propose a method of linear interpolation which overcomes any disadvantages of percentage / mark scaling and scaling based on cohort performance. We have also shared an easy-to-use excel file which shows you how to incorporate this method of scaling into your own assessments with ease.

Keywords: assessment, standard setting, scaling, health education.

Assessments in Higher Education are an important albeit challenging process; as the consensus statement from the Ottawa 2010 Conference stipulates, assessments need to be: valid in relation to their learning outcomes, consistent across cohorts, feasible, positive in influencing learning, supportive, and considered ‘acceptable’ to the students (Norcini et al., 2011). This requires careful and considered selection of assessment type (e.g. exam, practical, coursework; see Richardson, 2015), whilst following protocols and regulations set by the institution you teach at regarding timeframes, deadlines, and number of assessments. One of the key challenges however, arises when we start to consider practical skills – if we want to assess whether a student is competent in a particular skill, we need to work out how to get them to effectively demonstrate this in assessment conditions, and then decide what they need to showcase in order to pass – do they need to do the task perfectly, or just competently? This is a challenge for all educators, but particularly for programmes in healthcare, such as medicine, optometry, pharmacy, audiology, because it is also necessary to monitor and assess safe practice, skill competence, and professionalism throughout these practical assessments, all whilst managing a pass mark that aligns fairly with university-level criteria (40% for levels 4-6, 50% for stage 7+ in the UK; see Nash, 2023).

One way to help ensure validity in practical assessments is to utilise a process called ‘standard setting’ (see George et al., 2006; Yudkowsky et al., 2008; Tannerbaum and Kannan, 2015) which is recommended by professional bodies across the world (see BCRSP, n.d., College of

Optometrists, n.d., RCEM, 2022) to distinguish between competent and incompetent candidates. In practice, there are several recognised methods, including the Bookmark method (see Lewis, et al., 2011), the Angoff method (Angoff, 1971), and borderline regression (Kramer et al., 2003). However, generally speaking, in higher education standard setting would typically be criterion-referenced (absolute) where the cut score (pass mark) is determined by an expected level of performance. In other words, the cut score for each assessment would reflect the score expected of a ‘minimally competent’ candidate; the idea being that if a minimally competent candidate would achieve this score, then it is equivalent to the minimum pass mark for that assessment. Depending on the method, there are guidelines on best ways of deciding this, but the crucial factor is that it is no longer guaranteed that the pass mark for each assessment will be equivalent to the university-level pass mark, which, in instances where the assessment contributes a pass/fail component this is relatively easy to apply because students just need to demonstrate that they meet the criteria to progress. However, in cases where the component carries ‘weighting’, it can lead to differences in pass mark across different assessments within the same module, or the same programme, which can lead to difficulties in expectation management amongst students (i.e. “*why is 50% a pass mark this time but last time I needed 60%?*”), and can cause challenges with university software systems that anticipate pass marks to be 40% or 50% as appropriate. For example, if a hypothetical practical assessment has a cut score (pass mark) of 10/17, this would be calculated as equivalent to a student achieving 58.8% of the criteria. If a student achieved a fail score of 9/17 this may appear – to the student – that they have achieved 52.9% of the criteria and therefore should have passed the assessment according to the university boundaries.

A proposed solution then is to scale the marks to align with the university-level pass mark, within your institution’s regulations. Typically, this would include things like: students must maintain rank order, no scores below 0%, no scores above 100% etc. However, guidance on how to do this will often focus on changing the average grade of the cohort, or removing ‘problem’ questions from the analysis, which is inappropriate for standard setting where it is required for students to be scaled to a pass mark of 40%.

In our institution, we use linear interpolation which utilises linear polynomials to construct new data points from a known set of data points. In other words, it converts one range of numbers into another. The problem is, the goal is to scale anything above the cut score to be between 40%-100%, and anything below the cut score to be between 0%-39.9%, as appropriate, which requires two iterations of the same formula: one for passing marks and one for failing marks.

The formula itself for this is shown in Equation 1, and is the standard equation for re-scaling any range of numbers, where: $f(x)$ is the student’s scaled grade, x is the student’s criteria grade, and the remaining x and y values are either labelled as ‘old’ minimum (min) or maximum (max), or ‘new’ minimum and maximum as appropriate. All values should be input as percentages:

$$\text{Equation 1: } f(x) = y_{newMax} + \frac{y_{newMin} - y_{newMax}}{x_{oldMin} - x_{oldMax}} (x - x_{oldMax})$$

In the example above, the cut score for the hypothetical assessment is 10/17, meaning that 10/17 (58.8%) needs to be scaled to 40%. For this to work, the software of choice (e.g. Excel or MATLAB or equivalent) needs to know that if it is given a criteria score of 58.8% or above, then it scales between 40% (y_{newMin}) and 100% (y_{newMax}), whilst if it is given a criteria score of <58.8%, it should scale between 0% (y_{newMin}) and 40% (y_{newMax}). This can be achieved easily in Excel using IF functions as seen in Supplementary Material 1.

Figure 1 shows the original percentage for each possible score in the hypothetical assessment (red circles) alongside the scaled/ adjusted scores (blue circles) following the method described above.

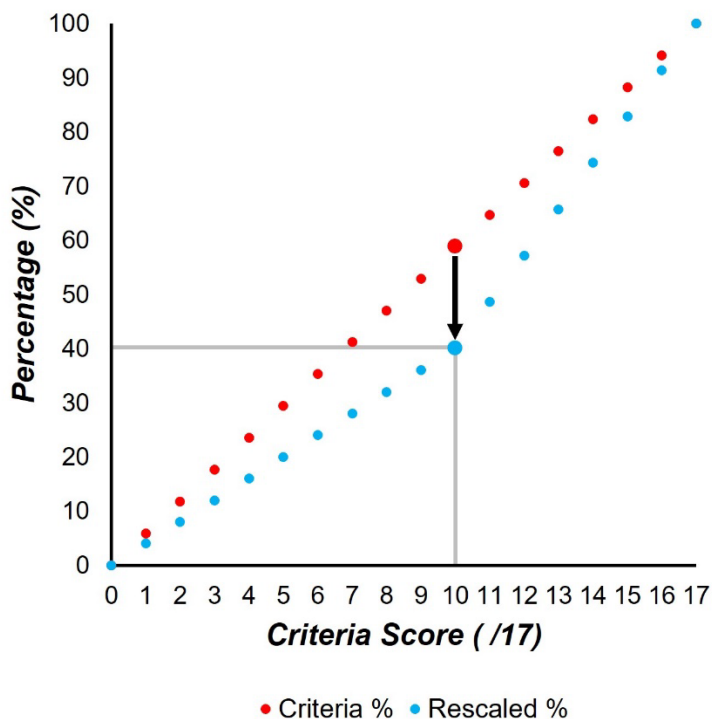


Figure 1. A plot showing hypothetical scaled data. These data are from a standard-set assessment marked out of 17 total criteria, with a cut score of 10/17. The original percentage of ‘criteria met’ are shown as red circles, whilst the scaled percentage are shown as blue circles. The larger circles connected by an arrow indicate the cut score (pass mark) that has been scaled down to 40% as appropriate. The grey lines show that this method converts the cut score to align with university-level grade boundaries.

There are many advantages of this method, namely that it adheres to common university regulations on scaling / adjusting marks, but it also protects very high scores (close to 100%) and very low scores (close to 0%), with no change at 0% or 100% and very small scaling for scores close to 0% or 100%. In this example, where the cut score is a lot higher than the university-level pass mark, scores close to the cut score will be scaled by the largest margin, where a cut score of 10/17 (larger circles) will be reduced by 18.8% to align with the pass mark of 40%. However, this is appropriate because the ‘criteria percentage’ is not a genuinely attained percentage in the university grading system, it is just a percentage of the criteria achieved successfully in that standard-set assessment, meaning it is *equivalent* to 40% at the undergraduate level.

Another advantage of this method is that it provides consistency across a programme and across cohorts because it aligns all pass marks. Alternative options for scaling allow for consideration of the distribution of grades determined by performance within a particular cohort, but those are not usually popular with students because a ‘well performing’ cohort might leave an individual student with a lower scaled grade than they might have achieved in a ‘poorer performing’ cohort. They are also more susceptible to bias when cohorts are small. With the method proposed above, the same equation can easily be applied for consecutive cohorts if the assessment is kept the same.

Overall, this is a neat and pragmatic solution to the problem of scaling standard-set marks in healthcare (practical competency) programmes within higher education.

References

- Angoff, W.H. (1971). The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. Retrieved January 24, 2024, from <https://files.eric.ed.gov/fulltext/ED050181.pdf>
- BCRSP. (n.d.). BCRSP Examinations. Retrieved November 30, 2023, from: <https://bcrsp.ca/en/prospective-certificants/bcrsp-examinations#:~:text=The%20pass%20mark%20is%20set,is%20the%20modified%2DAngoff%20method>
- College of Optometrists. (n.d.). Your OSCE Results. Retrieved November 30, 2023, from: <https://www.college-optometrists.org/qualifying/scheme-for-registration/final-assessment-osce/your-osce-results#:~:text=The%20Borderline%20Regression%20is%20a,OSCEs%20and%20is%20recognised%20internationally>
- George, S., Haque, M.S. and Oyeboade, F. (2006). Standard setting: comparison of two methods. *BMC medical education*, 6, 1-6. <https://doi.org/10.1186/1472-6920-6-46>
- Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L., & Van Der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, 37(2), 132-139. <https://doi.org/10.1046/j.1365-2923.2003.01429.x>
- Lewis, D.M., Mitzel, H.C., Mercado, R.L. and Schulz, E.M. (2011). 'The bookmark standard setting procedure', in Cizek, G.J. (ed.) *Setting performance standards* Routledge, pp. 225-253.
- Nash, K. (2023). How does the UK university grading system work? Retrieved January 23, 2024, from: <https://universitycompare.com/advice/student/uk-university-grading-system>
- Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M.J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V. and Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 206-214. <https://doi.org/10.3109/0142159x.2011.551559>
- RCEM. (2022). Setting the Pass Mark – The Angoff Method Explained. Retrieved November 30, 2023, from: https://rcem.ac.uk/wp-content/uploads/2022/03/Setting_the_Pass_Mark_Angoff_Method.pdf
- Richardson, J.T. (2015). Coursework versus examinations in end-of-module assessment: a literature review. *Assessment & Evaluation in Higher Education*, 40(3), 439-455. <https://doi.org/10.1080/02602938.2014.919628>
- Tannenbaum, R.J. and Kannan, P. (2015). Consistency of angoff-based standard-setting judgments: are item judgments and passing scores replicable across different panels of experts?. *Educational Assessment*, 20(1), 66-78. <https://doi.org/10.1080/10627197.2015.997619>
- Yudkowsky, R., Downing, S.M. and Popescu, M. (2008). Setting standards for performance tests: a pilot study of a three-level Angoff method. *Academic Medicine*, 83(10), S13-S16. <https://doi.org/10.1097/acm.0b013e318183c683>