In M. Bétrancourt, G. Ortoleva, & S. Billett (Eds.), *Writing for professional development* (pp.107–128). Leiden, The Netherlands: Brill.

McCrindle, A. R., & Christensen, C. A. (1995). The impact on learning journals on metacognitive and cognitive processes and learning performances. *Learning and Instructions, 5*(2), 167–185.

McGuire, L., Lay, K., & Peters, J. (2009). Pedagogy of reflective writing in professional education. *Journal of the Scholarship of Teaching and Learning, 9*(1), 93–107. Retrieved from http://files.eric.ed.gov/fulltext/EJ854881.pdf

McPeck, J. E. (1981). *Critical Thinking and Education*. Toronto: Oxford University Press. pp.v1, 170.

Mezirow, J. (1997). Transformative learning: Theory to practice. *New Directions for Adult and Continuing Education, 74*, 5–12.

Moon, J. A. (2013). *Reflection in learning and professional development: Theory and practice*. Oxon: Routledge.

Naber, J., & Wyatt, T. H. (2014). The effect of reflective writing interventions on the critical thinking skills and dispositions of baccalaureate nursing students. *Nurse Education Today, 34*(1), 67–72.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*(3), 259–271.

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117–175.

Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review, 16*(4), 385–407.

Pintrich, P. R., & Zusho, A. (2002). The development of academic self-regulation: The role of cognitive and motivational factors. In A. Wigfield, & J. S. Eccles (Eds.), *Development of Achievement Motivation* (pp.249–284). San Diego, CA, US: Academic Press.

Rodgers, C. (2002). Defining reflection: Another look at John Dewey and reflective thinking. *Teachers College Record, 104*(4), 842–866.

Ryan, M., & Ryan, M. (2013). Theorising a model for teaching and assessing reflective learning in higher education. *Higher Education Research & Development, 32*(2), 244–257.

Schunk, D. H. (2005). Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational Psychologist, 40*(2), 85–94.

Schunk, D. H., & Zimmerman, B. J. (1994). *Self-regulation of learning and performance: Issues and educational applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Scott, S. G. (2009). Enhancing reflection skills through learning portfolios: An Empirical Test. *Journal of Management Education, 34*(3), 430–457.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott, & J. Parziale (Eds.), *Microdevelopment: Transition Processes in Development and Learning* (pp.47–59).

Vygotsky, L. S. (1986). *Thought and language (Newly revised and edited by Alex Kozulin)*. Massachusetts: MIT.

Watkins, D., Dahlin, B., & Ekholm, M. (2005). Awareness of the backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lecturers. *Instructional Sciences, 33*(4), 283–309.

Whitebread, D. G., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., & Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning, 4*(1), 63–85.

Wilson, G. (2013). Evidencing reflective practice in social work education: Theoretical uncertainties and practical challenges. *British Journal of Social Work, 43*(1), 154–172.

Zimmerman, B. J. (2002). Achieving academic excellence: A self-regulatory perspective. In M. Ferrari (Ed.), *The pursuit of excellence through education* (pp.85–110). Mahwah, NJ: Erlbaum.

When can a case be made for using fixed pass marks?

Tom Bramley Research Division

Introduction

General Certificate of Secondary Education (GCSEs) and General Certificate of Education Advanced levels (A levels) have sophisticated procedures to ensure that the grade boundaries on examination components are set in places that achieve the goal of maintaining standards over time and between awarding organisations (AOs). Statistical methods currently have a prominent role. The 'comparable outcomes' method of The Office of Qualifications and Examinations Regulation (e.g., Ofqual, 2011; Benton, 2016) produces a target distribution of grades for each examination[1] and the AOs have to set boundaries on the components that result in an overall outcome that does not deviate beyond an allowed tolerance from these targets. Although there are good reasons for using these sophisticated procedures (including the prevention of 'grade inflation', and helping to ensure examinees are not disadvantaged when there is a major or minor system change), they do have drawbacks in terms of the resources required to administer them, both in staff time and in data availability. They are well-suited to the GCSE and A level case where there are only one or two examination sessions a year, large cohorts of examinees of roughly the same age are taking the exams, and large administrative data sets tracking the previous educational achievement of these examinees are available. However, some other high- and low- stakes assessment contexts do not have these advantages. In particular, many vocational and other non-academic assessments (such as the driving theory test) are either available on-demand or have multiple testing sessions, with widely fluctuating cohort sizes and groups of test-takers

1. The target distribution is for those examinees for whom there is a measure of prior attainment: Key Stage 2 score at GCSE, and mean GCSE score at A level.

from a wide range of ages, institutions and educational backgrounds. The AO or testing agency may have no information about the prior or concurrent achievement or ability of the group of test-takers and, in some cases, pre-testing is not possible because of cost or concerns about test security. Furthermore, in many such contexts the pass/fail (or other) decision needs to be made as soon as the test has been marked – and for computer-based tests this can be instantly, which requires the pass mark either to be known before the test is taken, or derivable from the items that were administered (in the cases where tests are compiled 'on-the-fly' or administered adaptively).

In some cases expert judgement can be used to arrive at a pass mark – for example by using a standard-setting method such as the Angoff or Bookmark methods (see Cizek, 2012, for a description of such methods). These methods often involve experts making judgements about the difficulty of test items, and the final decision can involve consideration of the potential impact on pass rates of setting the pass mark at particular scores. However, judgements of item difficulty can be unreliable and, as already noted, in some contexts the pass mark needs to be set before the impact on pass rates is known.

Using fixed pass marks, such as "To pass this test you need to answer 30 out of 40 items correctly" or "To pass this test you need to obtain more than 60 per cent of the available marks" might seem to be a simplistic solution to a complex problem. However, it does have some attractions, (Bramley, 2012), including:

- transparency: Test-takers know before taking the test how well they need to do in order to pass;
- validity of inferences about what test-takers know and can do. If past or example papers are publicly available then stakeholders can inspect these themselves and draw their own conclusions about the capability of someone who has achieved a given percentage of the marks available;
- perceived fairness for the test-taker: They know that their result did not depend on the performance of the other test-takers who happened to take the same test (or the prior attainment of other test-takers). However, this advantage could entirely disappear if different test forms are perceived to differ drastically in difficulty ('my friend got an easy set of questions');
- if the pass mark is fixed at a relatively high level then there is some reassurance that people who pass can actually answer most of the questions of the kind that were asked, which is important for 'consumer confidence' in some cases (e.g., a pass mark of only 50 per cent on knowledge of medical terms or routine procedures might not inspire confidence if it was part of a qualification for surgeons);
- the pass/fail decision can be made instantly (assuming the test is auto-marked); and
- the cost in money and staff time of setting the pass mark by more complex methods could be reduced.

The obvious drawback to using fixed pass marks is that it does not allow for the fact that test forms may vary in difficulty despite best efforts to construct or design them to be similar. The aims of the research described here were to investigate how serious a problem this might be in practice, and to explore the extent to which it could be alleviated by using expert judgement in the test construction process.

How much do tests randomly sampled from an item bank differ in difficulty?

A calibrated item bank[2] of 664 dichotomous items testing a single construct (Thinking Skills) classified into 7 different topic/skill areas was used as the basis for several simulations. The number of items and distribution of difficulties within each topic/skill area are shown in Table 1.

Table 1: Descriptive statistics for item bank (item difficulties in logits)

Topic/skill	*Total # items*	*Mean*	*SD*	*Min*	*Max*
1	122	0.39	0.99	-2.35	2.77
2	102	-0.08	1.10	-2.89	2.84
3	125	0.23	1.02	-2.17	3.64
4	120	0.42	1.15	-3.62	3.29
5	86	0.50	1.00	-2.72	3.73
6	57	0.01	1.13	-2.98	2.09
7	52	-0.10	0.75	-2.37	1.57
Total	**664**	**0.24**	**1.06**	**-3.62**	**3.73**

The simulated scenario was that a 40-item test with a fixed pass mark was to be constructed from this bank, with items from the different topic/skill areas represented according to their proportions in the bank[3]. The bank was 'recentred' by subtracting 0.24 logits from each item's difficulty to make the overall mean zero, and facilitate the interpretation of the minimum ability required to pass. This ability was arbitrarily selected to be 0.7 logits which, according to the Rasch model equation, corresponds to a probability of ≈0.67 of success on the average item. A thousand stratified random samples of 40 items were taken from the bank (stratified to ensure that the correct number of items testing each skill were included) and the 'correct' pass mark was calculated as the expected score that would be achieved by an examinee with an ability of 0.7 logits[4]. The average pass mark across all 1,000 tests was 25.6, so the nearest whole number value for a 'fixed' pass mark of 26 (65%) was taken, and compared with the correct pass mark (rounded to the nearest whole number) on each of the 1,000 tests.

Table 2 shows that 76% of the tests had a pass mark within 1 mark of the fixed pass mark of 26, and that 95% were within 2 marks. Figure 1 shows how the pass marks fluctuated from test to test.

One of the factors that affects how much pass marks fluctuate on tests constructed by sampling in this way is the underlying variability of difficulty in the whole bank. If all the items in the bank were the same difficulty, all tests constructed from it would be too. It is conceivable that different domains of knowledge/skill might differ in the extent to which test items might vary in difficulty. For example, if all the items require straightforward recall of basic factual knowledge gained on the course of study, there might be less reason to expect one item to differ too much from another in terms of difficulty. With that in mind, the entire bank was scaled by a factor of 0.8 to reduce the spread of difficulties and the process previously described was repeated.

2. The items were multiple-choice items calibrated using the Rasch model (e.g., Wright & Stone, 1979).
3. Specifically: 7, 6, 7, 7, 5, 4, 4 items from topic/skill areas 1–7 respectively.
4. This is the sum of expected scores on each item according to the Rasch model.

Table 2: Distribution of (absolute) differences from a fixed pass mark of 26 (full bank of 664 items)

FixedPassMarkDiff	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	319	31.90	319	31.90
1	446	44.60	765	76.50
2	187	18.70	952	95.20
3	40	4.00	992	99.20
4	8	0.80	1,000	100

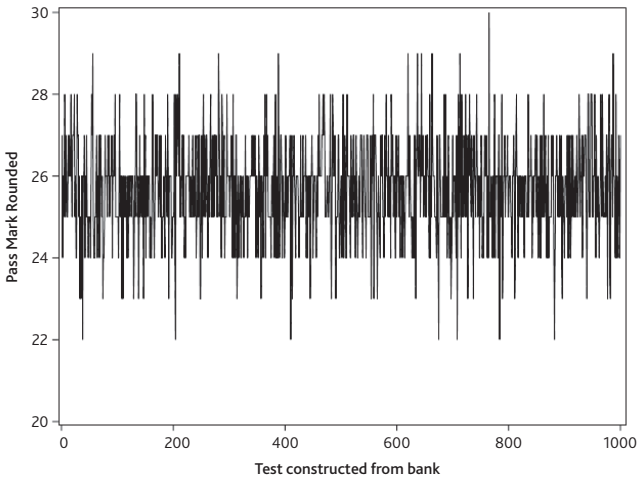

Figure 1: Correct pass marks on the 1,000 tests constructed from the full bank of 664 items

Table 3: Distribution of (absolute) differences from a fixed pass mark of 26 (664 item bank scaled by a factor of 0.8)

FixedPassMarkDiff	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	368	36.80	368	36.80
1	490	49.00	858	85.80
2	126	12.60	984	98.40
3	16	1.60	1,000	100

The scaling reduced the variability of the pass marks – over 85% of tests now had pass marks within 1 mark of the fixed pass mark of 26, and 98% of tests were within 2 marks. Of course, the scaling factor of 0.8 was entirely arbitrary, but this result shows that attempts to reduce the variability of item difficulty could contribute significantly to justifying using fixed pass marks.

The two simulations we have outlined used all the available calibrated items – 664 in total. In some testing contexts (e.g., the development of a new test) there may not be the luxury of such a large pool of items to draw from. A smaller bank of 200 items was therefore created by randomly sampling from topic/skill areas 1 to 6 according to the proportions (20%, 15%, 15%, 20%, 15%, 15%). The new smaller bank therefore had (40, 30, 30, 40, 30, 30) items representing these 6 topic/skill areas. Repeating the sampling process to construct 1,000 new tests gave a new (rounded) mean pass mark of 25, so this was now taken as the fixed pass mark. The resulting pass marks fluctuated in a very similar degree to those from the full bank.

In some contexts there may be rules or reasons preventing the sharing of items across test forms. For example, we could imagine that the 200 items in the smaller bank were constructed with the intention of creating 5 unique 40-item tests. It is therefore interesting to see how much pass marks would vary across sets of five tests (i.e., using every item in the bank) meeting the content specification but containing *no overlapping items*. A thousand such sets of five tests were constructed by random sampling as before (but without replacement). We are now interested in the extent to which the pass marks on each set of 5 tests differ from a set of 5 tests with a fixed pass mark of 25. One way to quantify this is simply to calculate the total absolute deviation across the 5 tests from the pass mark of 25. For example, a set of 5 tests with pass marks of (25, 26, 24, 24, 27) would score a total of 0+1+1+1+2 = 5.

Table 4: Distribution of total absolute deviation (TAD) from a pass mark of 25 across 5 non-overlapping tests in 1,000 sets of 5 tests constructed from the bank of 200 items and 6 topic/skill areas

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	10	1.00	10	1.00
1	24	2.40	34	3.40
2	102	10.20	136	13.60
3	113	11.30	249	24.90
4	231	23.10	480	48.00
5	149	14.90	629	62.90
6	165	16.50	794	79.40
7	104	10.40	898	89.80
8	61	6.10	959	95.90
9	23	2.30	982	98.20
10	16	1.60	998	99.80
11	2	0.20	1,000	100

Table 4 shows that nearly 63% of the sets had a total absolute deviation of 5 or less. A value of 5 would correspond to being 1 mark away from the fixed pass mark on all 5 (or to other combinations such as 2 above on 1, 3 below on another, and equal on 3). It was very rare (occurring only 1% of the time) for all 5 tests to have the fixed pass mark by chance.

Can expert judgement help to reduce the extent to which test forms differ in difficulty?

In the previous scenario 5 non-overlapping tests were constructed from a bank of 200 items. If the imagined scenario is adapted such that only four tests are needed operationally (with one as back-up for emergencies), then experts could be asked to identify, from the set of five, the four that appear most similar in difficulty (or, conversely,

the test that appears to be most different from the others in difficulty). Table 5 shows that when the most discrepant test from the 5 was removed (using the same data as in Table 4) then the percentage of sets of 4 with a total absolute deviation of 4 or less was nearly 86%, which compares well with the equivalent figure of 63% for the 5 tests. The percentage of sets where all 4 met the fixed pass mark was still low at 3.4%.

Table 5: Distribution of total absolute deviation (TAD) across best 4 non-overlapping tests from a pass mark of 25 in 1,000 sets of 5 tests constructed from the bank of 200 items and 6 topic/skill areas

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	34	3.40	34	3.40
1	136	13.60	170	17.00
2	223	22.30	393	39.30
3	277	27.70	670	67.00
4	189	18.90	859	85.90
5	85	8.50	944	94.40
6	39	3.90	983	98.30
7	16	1.60	999	99.90
8	1	0.10	1,000	100

Another way of capturing expert judgement of item difficulty might be to ask the item writers to rate individual items (e.g., as being of low-, medium- or high- difficulty. Would tests constructed to be of equal difficulty, in terms of the proportions of items in these three categories, be more likely to be of equal difficulty than tests constructed at random? In order to simulate expert ratings in three categories, a continuous variable was created to be correlated ≈0.7 with the item difficulties. (An average correlation of around 0.6 was reported in Brandon, 2004, between estimates of difficulty in Angoff-type standard-setting exercises and the empirical difficulty values). The top 50 items in the bank according to this variable were assigned a value of '3' (high); the next 75 items '2' (medium); and the bottom 75 items '1' (low). The correlations of this discrete variable with the actual difficulties turned out to be 0.64. This probably represents a slightly optimistic view about what might be achievable with expert judgement.

Figure 2 shows that there was some overlap in the three categories. Nevertheless there was a clear increase in difficulty with the judged category of difficulty. The next step was to construct sets of 5 non-overlapping tests from the bank that not only met the criteria of having the right number of items testing each topic/skill area, but also met the criteria for having the right number of items at each level of judged difficulty (i.e., 10 high, 15 medium, and 15 low). The algorithm written to do this started from a random selection (as before) but then within each test swapped items from over-represented levels of difficulty for items with under-represented levels of difficulty testing the same topic/skill area in the remaining pool of unselected items[5]. This took substantially more computer time to run, so 200 sets of 5 tests were created instead of 1,000.

5. The algorithm was not optimal (in many ways), one way being that the different skills were searched sequentially for items to swap. Thus, 'Skill 1' was always involved in any swapping and 'Skill 6' only very rarely.

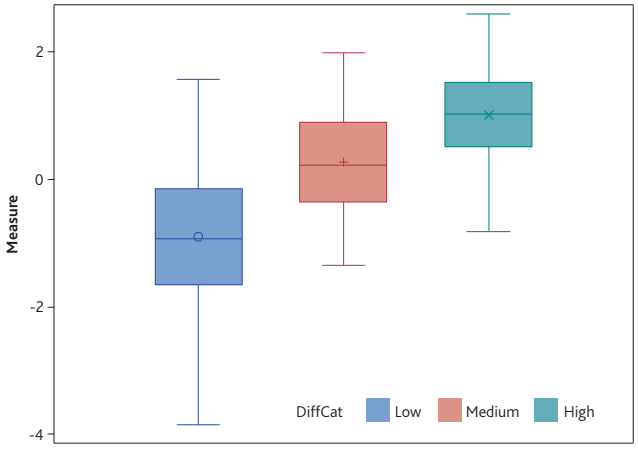

Figure 2: Relationship between 'judged' (simulated) difficulty category (DiffCat) and actual difficulty in the bank of 200 items

Table 6: Distribution of total absolute deviation (TAD) across five non-overlapping tests with the same distribution of judged difficulty

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	2	1.00	2	1.00
1	7	3.50	9	4.50
2	35	17.50	44	22.00
3	30	15.00	74	37.00
4	58	29.00	132	66.00
5	36	18.00	168	84.00
6	21	10.50	189	94.50
7	9	4.50	198	99.00
8	2	1.00	200	100

Comparing Table 6 with Table 4 shows that there was considerably less deviation of the pass marks. For example, 84% of the sets had a total absolute deviation of 5 or less compared with 63% using random selection.

Table 7 shows that if (after constructing 5 tests with the designated number of items at each level of judged difficulty) it were still possible for experts to identify the one furthest away from the average, then over 95% of sets of 4 would have a total absolute deviation of 4 or less (cf. 86% in Table 5).

Effect of overall ability distribution on fluctuations in pass rate

Finally, the effect on the pass rate of having fixed pass marks (as opposed to pass marks with the 'correct' value according to the bank difficulty) was investigated. The fluctuation in pass rate clearly is likely to depend on the ability (achievement/learning/knowledge) of the examinees in relation to the questions. When setting grade boundaries on A levels, there are usually relatively few examinees around the E boundary, and moving this boundary up or down by a few marks has little effect on the

Table 7: Distribution of total absolute deviation (TAD) across best four non-overlapping tests with the same distribution of judged difficulty

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	9	4.50	9	4.50
1	42	21.00	51	25.50
2	66	33.00	117	58.50
3	45	22.50	162	81.00
4	29	14.50	191	95.50
5	8	4.00	199	99.50
6	1	0.50	200	100

cumulative percentage of examinees achieving grade E. By contrast, there are usually many more examinees near the grade A boundary, and small changes in this boundary can have much larger effects on the cumulative percentage. To investigate the effect of the examinee ability distribution on pass rate fluctuation with fixed pass marks, a 'worst-case scenario' was simulated with a (normal) distribution of ability with a mean of 0.7 logits (i.e., around the pass mark, so 50% of examinees would be expected on average to pass the test) and standard deviation (SD) of 1 logit. Then this distribution was shifted by adding a constant amount such that around 80% of examinees would be expected to pass the test. The scores of 1,000 (different) examinees on each of the (randomly constructed) 1,000 tests from the 200-item bank were simulated using the Rasch model. Figure 3 shows the simulated score distributions for the first of these 1,000 tests.

Table 8: Descriptive statistics for distributions of simulated pass rates

		N	*Mean*	*SD*	*Min*	*Max*
Mean at pass mark	True	1,000	52.44	2.18	45.7	58.3
	Fixed	1,000	52.64	6.42	29.9	69.9
Mean above pass mark	True	1,000	80.31	1.65	75.7	84.6
	Fixed	1,000	80.25	4.54	59.9	90.6

Figure 4 and Table 8 show that there is considerably more variability (SD ≈3 times greater) in pass rates using fixed pass marks from tests constructed at random than from using the correct (true) pass marks, but that, as expected, the variability (and the difference between true and fixed) is less when the bulk of the distribution is some distance away from the pass mark.

Summary and possible rationale for using fixed pass marks

In summary, the simulations have shown that:

- tests constructed by random sampling from an item bank vary in difficulty;
- with a pass mark at around 60–65% of the maximum mark, around 75% of 40-item tests constructed at random from the particular real item bank used as a basis for this work would have a pass mark within 1 mark of the fixed pass mark;
- this percentage would be greater if the items in the bank had a lower spread of difficulty (and vice versa);
- constructing 5 non-overlapping tests (i.e., with no items in common) at random from a bank of 200 items produced around 63% of sets of 5 where the total absolute deviation from fixed pass marks was 5 or less (i.e., an average discrepancy of 1 mark per test);
- this could be increased (to 86%) for 4 tests if experts could infallibly identify the most discrepant test in a set of 4;
- constructing 5 non-overlapping tests to meet criteria of equal difficulty as defined by expert judgement (assumed to correlate around 0.6 with actual difficulty) produced around 84% of sets of 5 where the total absolute deviation from fixed pass marks was 5 or less; and
- the variability in pass rates from tests with fixed pass marks is around three times greater than from tests with the correct pass mark, but the amount of variability (for both) depends on where the distribution of examinee ability is in relation to the pass mark. If the average pass rate is around 80%, the variability (SD) in pass rate is around three-quarters of what it is if the average pass rate is around 50%.

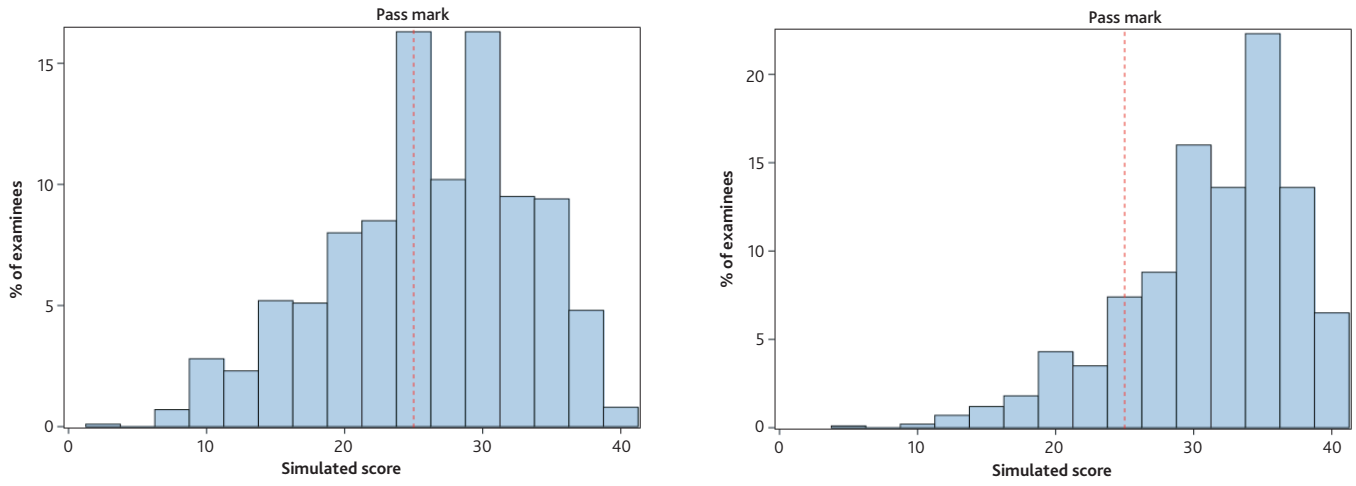

Figure 3: Simulated score distributions on one test for cohort with mean ability around the pass mark (left) and with mean ability above the pass mark (right)

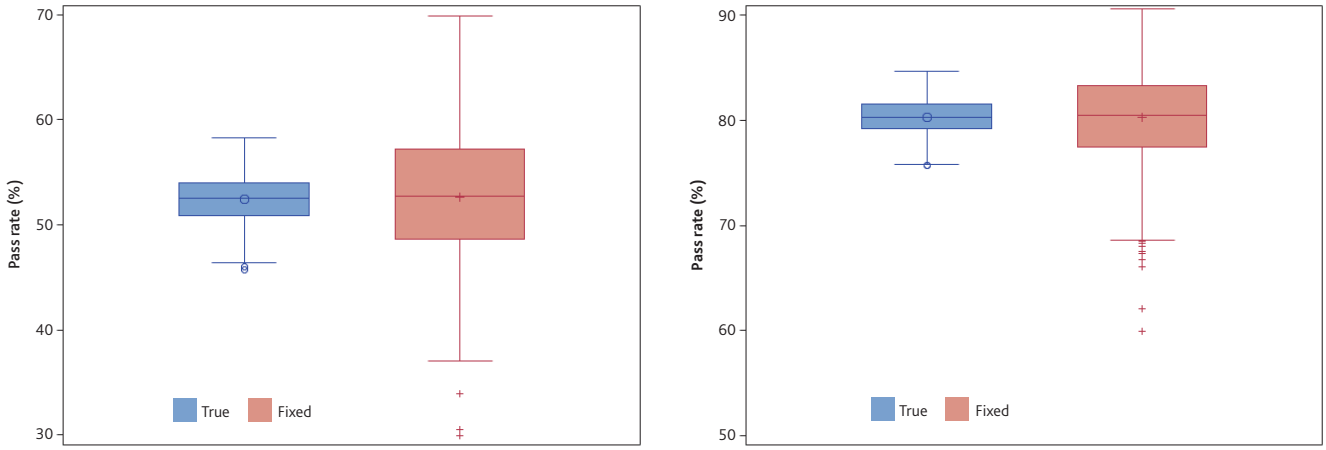

Figure 4: Distribution of pass rate using the true or fixed pass mark for cohorts with mean ability around the pass mark (left) and with mean ability above the pass mark (right)

One possible rationale for using fixed pass marks would be to conceive of the knowledge domain in each subject as a finite set of questions that could possibly be asked. We want to infer that the proportion of the domain known by examinees is above a certain value (e.g., 70%). If a test is constructed by stratified random sampling from the domain, then the proportion they get right is an unbiased estimate of the proportion of the domain that they know. The pass mark on the test could be set at the same percentage as the target domain percentage (i.e., 70%) or it could be adjusted to allow for the cost of making a false positive or false negative error (e.g., if it were deemed more costly to fail someone who knew more than 70% of the domain than it would be to pass someone who knew less than 70%).

The main challenge to this idea would be that individual tests would still differ in difficulty and it would be unfair to examinees not to try to allow for this somehow (as currently happens in GCSEs and A levels). This is of course a good point, but there are some possible responses. Firstly, we could argue that factual knowledge does not fit the concept of a 'latent trait' in the way that, for example, mathematical ability does. That is, there is arguably no real concept of a hierarchy of item difficulty that could define a meaningful continuum of progression. That is not to say that some items will not be answered correctly by more people than other items, but that the factors that make particular items of knowledge 'easy' or 'difficult' to recall will be idiosyncratic to the particulars of the learning experience and interests of different individuals. Tests of factual knowledge are therefore, in a sense, by definition equally difficult.

Secondly, when numbers of examinees are low, attempting to equate tests by statistical methods (e.g., comparable outcomes) can introduce more random error than it removes systematic error. In GCSEs and A levels, the grade boundaries on examination components are often unchanged when very few examinees have taken the component.

Thirdly, in an on-demand testing context (e.g., a test which is computer-delivered and auto-marked) when tests are constructed from a bank such that different individuals take different tests, statistical definitions of equivalent scores based on the performance of large groups could be less relevant. A given individual might have a better chance of passing on Test A than Test B, even if in a large group more would pass B than A.

Fourthly, being a victim of bad luck is not quite the same as being a victim of unfairness. If an individual happens to receive a selection of items that they do badly on when they would have done better on other possible selections of items, this is bad luck for them. The potential for unfairness perhaps resides more in how costly (in terms of time, money, and missed opportunities) it is for the individual to re-take the test.

Finally, this research has shown that there are steps that can be taken to reduce the amount by which different tests fluctuate in difficulty – such as trying to reduce the range of item difficulty, and making use in the test construction process of any information we have in advance about item difficulty, such as expert judgements. In testing contexts where the reuse of items is permitted, accurate empirical data will over time replace the more fallible expert judgements and allow test forms of equivalent difficulty, and hence the same pass marks, to be created with increasing precision.

Returning to the question posed in the title of this article, people will differ in the weight they give to different considerations when reaching a judgement. In my opinion, for on-demand tests that mainly require recall of facts in well-defined domains, with groups of test takers that vary in size and demographic composition, the advantages slightly outweigh the disadvantages.

References

Benton, T. (2016). *Comparable Outcomes: Scourge or Scapegoat?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/346267-comparable-outcomes-scourge-or-scapegoat-.pdf

Bramley, T. (2012, February). *What if the grade boundaries on all A level examinations were set at a fixed proportion of the total mark?* Paper presented at the Maintaining Examination Standards seminar, London. Retrieved from http://cambridgeassessment.org.uk/Images/459357-what-if-the-grade-boundaries-on-all-a-level-examinations-were-set-at-a-fixed-proportion-of-the-total-mark-.pdf

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Cizek, G. J. (Ed.) (2012). *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd Ed.). New York and London: Routledge.

Ofqual (2011). *Maintaining standards in GCSEs and A levels in summer 2011.* Retrieved from http://webarchive.nationalarchives.gov.uk/20110718105952/http://www.ofqual.gov.uk/files/2011-05-16-maintaining-standards-gcses-and-alevels-summer-2011.pdf

Wright, B. D., & Stone, M. (1979). *Best Test Design*. Chicago, Illinois: MESA Press.