The Application of Simple, Practical and Affordable Standard Setting Methods in Small Groups of Students

John David Wright[1]

Corresponding author: Dr John Wright john.wright@mopipi.ub.bw
Institution: 1. Faculty of Medicine, University of Botswana
Categories: Assessment

Received: 13/10/2016
Published: 17/10/2016

Abstract

Background

A simple method of examination standard setting, the Cohen method, and a modification of it have been previously described, in which the exam pass mark is taken as a fixed percentage of the mark obtained by near-to-top performing students. These methods have proved effective in large cohorts, but their application to smaller groups has not been documented.

Aims

To retrospectively apply the two methods to the exams and tests of eight cohorts of around 50 students from years 1 and 2 of the MBBS course of a new African medical school.

Methods

The standard setting methods were applied to historical data from 127 tests and exams. Failure rates resulting from their use were compared to the failure rates resulting from the fixed pass mark of 50%.

Cumulative distribution function charts (CDFs) were obtained from all tests to assess the validity of the reference points used to calculate pass marks.

CDFs were used to assess the power of tests to discriminate between students. The performance of Cohen – type standardisation methods in tests of different discriminating ability is analysed and discussed.

Results
Both methods reduced the average failure rates resulting from tests and greatly improved the consistency of failure rates. The methods do not produce valid pass marks in tests which discriminate poorly.

**Conclusion**

The simple Cohen and modified Cohen standard setting methods can be used successfully in small groups. The methods produce fairer results than fixed pass marks or norm referenced methods but may produce anomalous results in tests which are extremely discriminating or which discriminate poorly.

**Keywords:** Standard setting, Cohen method, Small groups, Assessment

**Introduction**

It is axiomatic that medical examinations and tests which contribute to progression through a programme or to final qualification should undergo some form of standard setting. The aim of standard setting procedures is essentially to fix the pass mark of the assessment at the mark which would be obtained by a minimally acceptable candidate ie. a student who is safe to practice or who's knowledge and understanding is sufficient to allow them to cope with the next stage of the course. In a medical course with the usual highly selective admission requirements it would be expected that most candidates would meet this standard.

Though the aim of standard setting is simple to define, to actually meet that aim is extraordinarily difficult, and it is accepted that there is no ideal way to do it (Downing, Tekian and Yudkowski, 2006; Norcini, 2003). Fixed pass marks and norm referenced methods have often been used to set pass marks (Cohen-Schotanus and Van der Vleuten, 2010). The use of a fixed pass mark produces large variations in failure rates when the difficulty of examinations inevitably varies. Norm referenced methods produce consistent failure rates. For example, setting the pass mark at the mean mark of the group minus one standard deviation produces a constant failure rate of around 17%, as statistically it must. Since no attempt is made to reference this pass mark to any criteria these methods run the risk of failing acceptable candidates or passing unacceptable ones.

It is generally perceived by medical educators that standard setting methods should be criterion referenced (Banderanayake 2008). Several different criterion referenced methods have been described (Downing et al. 2006; Norcini, 2003), but there is no recognised ‘gold standard’. The methods described by Angoff (1971) Ebel (1979), and Nedelsky all use a panel of experts to set the pass mark. This involves gathering together around 15 doctors and medical educators with mixed backgrounds, who meet to formulate the aims of the procedure, then consider the exam paper question by question. The Hofstee method (Hofstee 1983) similarly uses a panel of experts but panel deliberations are less complex and the setting of the final pass mark involves an element of norm referencing, which makes the method less acceptable to some (Cohen-Schotanus and Van der Vleuten 2010). While these panel methods are intellectually satisfying, and are regarded as the best available procedures (Cohen-Schotanus and Van der Vleuten 2010; Taylor 2011) against which other methods may be judged, they are difficult to organise, time consuming and extremely expensive. This means that it is only feasible to use them for ‘high stakes’ exams. Moreover it has been shown that different panels reach quite different pass marks from the same set of questions (Boursicot, Roberts and Pell, 2006).

Most medical degree programmes contain multiple tests and exams in each year of the course, which, though they may not individually and alone determine qualification or progression, nevertheless contribute to progression decisions. These ‘low stakes’ tests and exams are too numerous to have panel based standardisation procedures...
applied to them, so they have usually been assessed using a fixed pass mark or a norm referenced method. It is desirable to standardise these tests in some way, since failure may often delay student's progress through the course. Recently two simple and practical standard setting procedures have been described (Cohen-Schotanus and Van der Vleuten 2010, Taylor 2011) which allow standard setting in these 'lower stakes' exams. These methods use the marks obtained by the best performing students as a reference point which defines the difficulty of the exam. The pass mark is fixed as a pre-set percentage of this reference point mark. In order to avoid the occasional off scale genius distorting the pass mark Cohen-Schotanus and Van der Vleuten (2010) use the 95\textsuperscript{th} percentile student mark as the reference point, setting the pass mark as 60\% of this, after allowing for randomly obtained marks. This can be expressed by the formula:

$$\text{Pass Mark} = R + 0.6(X - R)$$

where $R$ is the mark which could be obtained by random guessing, and $X$ is the mark of the 95\textsuperscript{th} percentile student.

This procedure is now usually called the Cohen method.

Taylor (2011) found the mark of the 90\textsuperscript{th} percentile student to be more reliable in her programme, and sets the pass mark as 65\% of this, making no adjustment for random marks. This reduces the formula to:

$$\text{Pass Mark} = 0.65Y,$$

where $Y$ is the mark of the 90\textsuperscript{th} percentile student.

This method is usually referred to as the modified Cohen method.

While these methods are not truly criterion referenced, since the examination questions are not considered when the pass mark is set, the use of marks obtained by students who took the test to set the pass mark ensures that test difficulty is taken into account.

Phase 1 (years 1 and 2) of the University of Botswana MBBS programme consists of 15 system-linked blocks most of which are examined by an end of block test plus an end of semester exam. These together make up 80\% of the mark for the block. There is also an anatomy exam and an OSCE in each year. Since the start of the programme in 2009 a fixed pass mark of 50\% has been used in all tests and exams, in line with university procedures in other programmes. However this has given problems with high failure rates from time to time when a test has proved to be unusually difficult and since all blocks have to be passed for a student to progress to the next year it is desirable that all tests and exams are standard set. Thus the simple standard setting methods described above had immediate appeal since there was no possibility of using panel methods.

Both methods described use marks below the very best in each year group to define their set reference point, because they are able to show, using the cumulative distribution function from their sets of exam marks, that the very top students often produce scores out of line with the rest of the group, so that using the best mark as the reference point will inflate the pass mark. This was why Taylor (2011) chose to use the 90\textsuperscript{th} centile rather than the 95\textsuperscript{th} centile mark. Since the methods were applied to large groups of students (Taylor’s group consisted of around 370 students. Cohen-Schotanus and Van der Vleuten 2010, do not state the group size but they are derived from two large medical schools) they could be reasonably certain that a 90\textsuperscript{th} or 95\textsuperscript{th} percentile student will be very able but not beyond the normal ability range so that they distort the pass mark. The group size in the University of Botswana School of Medicine is around 50 students per year (42 – 54 in the year groups analysed) so does not give the same statistical confidence. In a group of this size the 95\textsuperscript{th} percentile student is the third highest mark, and the 90\textsuperscript{th} percentile is the fifth highest, and since entry to the programme is very competitive, it seems possible that several high end students could be well outside the normal range of ability, in some groups at least.
A major aim of this study was to determine how the Cohen and modified Cohen standardisation methods performed when applied to University of Botswana School of Medicine Years 1 and 2 tests and exams, and in particular to examine whether the methods could be reliably applied to groups of only 50 students.

An additional aim was to use the cumulative distribution function (CDF), which was plotted for every test, to determine the reliability of the 95th and 90th percentile marks as reference points for the determination of the pass mark.

It will be apparent that the Cohen and modified Cohen methods require that a consistent relationship exists between the reference point mark (95th or 90th percentile candidate) and the mark of the minimally acceptable candidate. This will only be true if tests are equally discriminating between able and less able students. The discriminating power of a test can be estimated by measuring the gradient of the CDF for each test. Thus the discrimination of tests was estimated and the performance of the two standard setting methods in tests with different levels of discrimination was investigated.

**Methods**

The results from all tests and exams in years 1 and 2 taken in the four academic years 2012/13 to 2015/16, (127 tests in all), have been analysed. Pass marks for each test were calculated using the Cohen and modified Cohen methods and the number of fails which would have resulted determined and compared to the number resulting from use of the fixed pass mark of 50%. Fail rates higher than 15% were specifically noted since it has been suggested that this is the highest level of fail rate that could reasonably be expected given the admission requirements of the programme (Prof. John Cookson, Emeritus Dean, Hull and York Medical School, UK, Personal Communication).

In order to assess whether the 90th and 95th percentile students are responding to the test in the same way as the rest of the group, and to measure the discrimination ability of each test, a cumulative distribution function (CDF) was plotted for each of the 127 test results. The CDF is simply a plot of the rank of each candidate against the mark they obtained, but the candidate's rank is converted to a percentile in the group. This allows gradients of CDFs to be directly compared irrespective of the number of candidates taking the test. This gradient, which shows the percentile increase per mark, measures the test's ability to discriminate between candidates. The possible forms and features of the CDF are shown in fig 1.
The straight line of line A shows that the test is discriminating evenly across the whole of the student group, and any reference point chosen should be valid.

Although many tests give CDFs like line A, in many cases there is an upper inflection point, as shown in dotted line B. This shows that the top end students are performing proportionally better on the test than the others. If the chosen reference point (90th or 95th centile) falls in this region it will result in an elevated pass mark.

Dotted line C shows an inflection in the lower part of the CDF. This is produced by a number of students who are performing below the general level of the group, and may be characteristic of some groups.

Line D is parallel to line A, but at a lower mark level. This test will fail many students if a fixed pass mark is used, but it discriminates similarly to test A. Students have found this test difficult but it has discriminated well between them. This test is an excellent candidate for Cohen type standard setting.

Line E has a steeper gradient than line A, showing that it discriminates less well between candidates. This test will produce few failures using a fixed pass mark or a Cohen type standard set pass mark. It is likely to produce ‘false positives’.

For the purposes of this analysis inflections in the upper part of the CDF as exemplified by line B are important since
they show the possibility of an invalid reference point. The percentile position of this inflection point was noted on each CDF in which it occurred.

**Ethical Review**

University of Botswana has federal wide association (FWA) registration with the United States National Institute of Health for ethics purposes. Under NIH codes CFR 46 educational research aimed at quality improvement is exempt from ethical review.

**Results and Discussion**

Table 1 compares the numbers of students failed when a fixed 50% pass mark is used, with the failures which would have resulted if the Cohen or Modified Cohen method had been used on the results of the same tests.

<table>
<thead>
<tr>
<th>Student Cohort</th>
<th>50% pass mark</th>
<th>Cohen-set pass mark</th>
<th>Modified Cohen-set pass mark</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Av. fails %</td>
<td>SD %</td>
<td>Fail Range%</td>
</tr>
<tr>
<td>Yr 1 12/13</td>
<td>13.8</td>
<td>10.6</td>
<td>2-34</td>
</tr>
<tr>
<td>Yr 1 13/14</td>
<td>11.5</td>
<td>15.1</td>
<td>0-47</td>
</tr>
<tr>
<td>Yr 1 14/15</td>
<td>7.2</td>
<td>7.6</td>
<td>0-28</td>
</tr>
<tr>
<td>Yr 1 15/16</td>
<td>11.8</td>
<td>9.6</td>
<td>0-34</td>
</tr>
<tr>
<td>Yr 2 12/13</td>
<td>22.7</td>
<td>17.5</td>
<td>2-60</td>
</tr>
<tr>
<td>Yr 2 13/14</td>
<td>7.3</td>
<td>10.2</td>
<td>0-38</td>
</tr>
<tr>
<td>Yr 2 14/15</td>
<td>7.2</td>
<td>5.4</td>
<td>2-20</td>
</tr>
<tr>
<td>Yr 2 15/16</td>
<td>10.2</td>
<td>10</td>
<td>0-44</td>
</tr>
</tbody>
</table>

The table shows that both standard setting methods give average failure rates which are lower than that produced by the fixed 50% pass mark in all eight year groups analysed. However, what is more important is that the standard deviation of the percentage failures is also much lower in each group when either standard setting method is used. This shows that the standard setting methods give a much more consistent failure rate than the fixed pass mark.
Table 2 analyses all the 127 tests. It shows the number of tests with failure rates over the 15% which could be regarded as the maximum acceptable failure rate for these groups. Both methods dramatically reduced the number of tests which produced failure rates in excess of 15%, and limited the fail rates even in those tests which did exceed this figure. Analysis of raw test results shows that in many tests the standard setting methods produced pass marks close to 50%, and gave similar fail rates to the fixed pass mark. They produced lower pass marks, and hence lower fail rates in the few exams which students found particularly difficult, thus eliminating, or at least reducing, the excessive fail rates produced by these exams. They also often produced pass marks higher than 50% in exams which students found particularly easy, and sometimes introduced a small fail rate into these tests, when the fixed pass mark produced a zero fail rate. This also adds to the improved consistency achieved by the standard setting methods.

Table 2 also shows the range of pass marks which the standard setting methods produce, since this can have a bearing on the acceptability of the use of the method to the academic staff setting the test. Both methods produce a majority of pass marks in the 45 – 50% range. When the two standard setting methods are compared it is found that the Modified Cohen method usually produced lower pass marks than the original Cohen method, in spite of its use of a higher multiplier. The reason for this is not that its fixed point is much lower, but that a correction for random marks is not used. In fact, in tests such as OSCEs and anatomy spotter tests in which there is no random mark the modified Cohen method produced higher pass marks than the original Cohen method, and these accounted for most of the higher fail rates which were generated by this method.

To check out the possibility of groups of top end students responding to tests differently to the main part of the group, cumulative distribution function charts were produced for each of the 127 tests and exams. The possible forms of the CDF charts are shown in Fig 1 and discussed in the methods section. To answer the question about the possibility of a high performing top end group, the position of the start of the upper inflection in a CDF chart of the type shown in line B of Fig 1 was measured. Table 3 shows the results obtained, set out in year groups since the shape of the CDF chart might be expected to be a characteristic of a particular cohort of students.
Table 3 shows that although there are a significant number of tests in which groups perform homogeneously (ie. the CDF has no upper inflection point) in fact most CDFs had an upper inflection point, showing that usually some top end students responded to the test differently to the group as a whole. In most cases the inflection point occurred between the 95th and 99th percentile, which in a group of 50 students means that only the top one or two students responded in this way, and in these cases the 95th percentile mark would be safe to use as a fixed point. However, it does seem that some year groups may have a bigger group of ‘out of line’ top end students (in Table 3, Year 2 2014/15 is the best example) where the inflection point occurs between the 90th and 95th percentile, or even in some cases around the 85th percentile, in a significant proportion of the tests. In these cases it is possible that the fixed point used for standard setting might be too high, particularly for the original Cohen method.

In these latter cases it is possible to correct the fixed point, using the CDF chart, if the straight line part of the chart is projected forward to intersect with the 90th and 95th percentile line. This procedure has been applied to the tests from the Year 2 2014/15 group. The CDF plots of the results of the tests from this cohort were examined to find the ones with upper inflection points below the 95th percentile. The correction procedure described was used to find a corrected 95th and 90th percentile mark for these tests, and this was used to recalculate the Cohen and Modified Cohen pass marks. Just two tests were found in which the recalculated pass marks were significantly different to the original marks. In one the original Cohen set pass mark decreased from 49.8% to 45.4% and the modified Cohen pass mark went from 48.2% to 43.9%. This resulted in the fail rate for the Cohen standardised test falling from 6% to 4%, and for the modified Cohen standardised test falling from 4% to 2%. The failure rate for the fixed 50% pass mark was 6% for this test. In the second test the corrected Cohen pass mark fell from 48.1% to 44.8%, reducing the fail rate from 10% to 6%, while the corrected modified Cohen pass mark reduced from 46.5% to 43.6%, reducing the fail rate from 8% to 4%. The fixed 50% pass mark resulted in a 16% fail rate for this test.

This correction procedure illustrates just one way in which the cumulative distribution function charts from test results can be used to give valuable graphic information about the performance of students on the test. As already noted CDF charts with an upper inflection point show a group of top end students performing significantly better than the group generally. The percentile point at which this inflection occurs indicates the size of this group. Similarly an inflection point in the lower part of the chart (as in line C of Fig 1) can identify a group of poorly performing students.

The gradient of the CDF plot is an indicator of the degree of discrimination of the test, with a steep line showing a narrow spread of results and a test which discriminates poorly.
Figure 2 shows the CDF gradient measured for each test, plotted against the number of failures of that test when its pass mark was determined by the Cohen method (a plot using failures generated by the modified Cohen method was very similar, so is not shown). As expected, failures reduce as the CDF gradient increases ie, as the tests become less discriminating, but only up to a gradient of around four percentile points per mark. At gradients greater than this tests are only capable of picking out students performing well below the general level of the group. The fixed point standard setting methods being examined here are thus likely to produce 'false positives' (ie, to pass inadequate students) in poorly discriminating tests, or alternatively produce 'false negatives' (ie. to fail adequate students) in tests which discriminate markedly. It might be expected that calculation of an average mark and standard deviation would give the similar information, but these are easily distorted by a few poorly performing students, and the CDF chart is more reliable. Production of a CDF chart of test results is very simple in EXCEL, and it is strongly recommended.

The conclusion drawn from these results is that the use of a fixed point pass mark cannot be justified if useable alternatives are available. Both standard setting methods examined here performed well. Their merit is that they take into account the difficulty of the exam when fixing the pass mark, lowering the pass mark when students find the exam difficult and raising it when they find the exam easy. When introducing a standard setting method for general use this latter point is important since the method must be acceptable to teaching faculty, and a method which sometimes raises the pass mark above the one they are used to when the exam is easy will find readier acceptance than one which always lowers the pass mark.

The question as to which of the two methods works best in the local situation is more difficult. On the face of it the modified Cohen method has advantages. It produced generally slightly more consistent fail rates than the original Cohen method, (Table 1) and the data presented here agree with those of Taylor (2011) that the 90th percentile student may be a better reference point than the 95th percentile (Table 3). However, some faculty find not using a
correction for guessing difficult to accept. While it is accepted that applying a correction for guessing disadvantages certain personality types (Betts, Elder, Hartley and Trueman, 2009), this applies only when students are taking the exam. No correction for guessing is applied in the examinations themselves. When standard setting is done, correction for guessing takes place after the examination and students are not aware of it, so the important psychological considerations outlined by Betts et al.(2009) do not apply. The method used might gain easier acceptability with faculty if correction for guessing is done and there seems to be no valid reason for not doing it. The suggested local procedure therefore is to use original Cohen method with the 95th percentile mark as the reference point, but to plot CDF charts to check the validity of the 95th percentile mark, correcting it on the rare occasions when that is necessary. It should be noted that it is very easy to adjust these methods to suit local conditions, by changing the reference point and/or the multiplier used, or by using or not using correction for guessing, but this should only be done in the light of evidence.

Though the standard setting methods produce far more consistent failure rates than fixed pass mark methods there is no evidence at all that the pass level they set represents the level of the minimally acceptable candidate. Taylor determined the multiplier (0.65) used in the modified Cohen method by using the marks of exams set by the Angoff method, but was not able to do further checks against Angoff set exams. It has similarly not been possible to do that in this study because there are no Angoff set exams to check against.

The methods are also not ideal in other ways. In particular different levels of discrimination in tests will affect failure rates, with poorly discriminating tests tending to give false positives ie. pass unacceptable candidates, whilst highly discriminating tests may give false negatives ie. fail acceptable ones. The degree of test discrimination will be apparent in the gradient of the CDF chart, but it is not easy to modify the multipliers to take this into account.

Both Cohen-Schotanus and Van der Vleuten (2010) and Taylor (2011) regard the Angoff panel method as the best available, because the panel does examine all questions and attempts to decide how a minimally acceptable candidate could be expected to answer them. However it has been shown experimentally that five different Angoff panels produced quite different pass marks when standard setting the same examination (Boursicot et al. 2006), showing that the panel procedure is quite fallible. In fact the best test of validity would probably be a cohort study of borderline candidates, following them to the next stage of the programme or into their internship programmes as appropriate, to determine if they are in fact acceptable at the next level, or as a qualified junior doctor. Even this procedure, while it might show that the candidates who passed were acceptable, could not show anything about the candidates who were failed.

**Conclusion**

The standard setting methods described by Cohen-Schotanus and Van der Vleuten (2010) and modified by Taylor (2011), which set the pass marks of exams and tests as a fixed proportion of the mark of top performing students, are simple to use and quick to apply. Both the original method and its modification proved to be highly effective in reducing the variability in failure rates when applied to the examinations and tests taken over a four year period in the first two years of the MBBS programme of a new African medical school with cohort sizes ranging from 42 – 54 students.

Because of the small cohort size there had been concerns about the validity of using the marks of the 95th percentile student (original method) or 90th percentile student (modified method) as the reference point to set the exam pass mark because these are much nearer to the top mark of the group in a small cohort than they would be in a bigger group, and Cohen-Schotanus and Van der Vleuten (2010) showed that the top mark in the group was not the best
reference point to use. Using the cumulative distribution function charts of test results, however, has shown that both
the 95th and 90th percentile marks are usually valid. Because the small group size can sometimes lead to unusual
group behaviour it is recommended that cumulative distribution frequency charts be plotted for all sets of test results
in small groups to check that nothing unusual is happening, and correct for it if it is.

The use of CDF charts in this study has graphically illustrated the need for a certain level of discrimination between
candidates within tests if this type of standard setting method is to be used. This will apply to big groups as well as
small ones. This illustrates the importance of taking item analysis into account when constructing tests in order to
produce tests with adequate discrimination.

Though panel methods such as Angoff (1971) provide the security of a method which is demonstrably the best that
can be done, they are so time consuming and expensive that they cannot be widely used. The simpler methods tested
here provide standard setting methods which can be used on any test, and can be applied to the many small tests
which often make up large parts of the assessment of a programme. Both the original method and its modification
performed well in this study and there was little to choose between them. In fact the choice between them will
probably depend on whether or not the School prefers to use a correction for random marks when standardising their
tests. In spite of some deficiencies in the methods, which are discussed in this paper, they are far better than any
alternative. The use of one or other of the methods, or a similar method with a locally calculated multiplier, is
recommended in place of a fixed and essentially random pass mark.

Take Home Messages

The Cohen and Modified Cohen methods of standard setting are simple, practical and affordable for low stakes
assessments and can be applied just as successfully in groups of around 50 students as in larger cohorts.

Simple standard setting methods greatly reduce the variation of failure rates in tests compared to those produced by
the commonly used fixed pass mark.

The Cohen approach to standard setting can be readily customised by adjusting the reference point and/or multiplier
used, and using or not using correction for guessing, if local evidence warrants it.

A cumulative differential function plot of test results gives valuable information about test performance, including
validation of the Cohen and Modified Cohen reference points.

Low or high discrimination in examinations or tests can give ‘false positives’ or ‘false negatives’ using these methods.

Notes On Contributors

John D. Wright BSc PhD is Senior Lecturer in the Department of Biomedical Science and an Academic Associate
of the Department of Medical Education in the Faculty of Medicine, University of Botswana.

Acknowledgements
Bibliography/References


http://dx.doi.org/10.1080/01421590802402247


http://dx.doi.org/10.1080/02602930701773091


http://dx.doi.org/10.1007/s10459-005-5291-8

Cohen-Schotanus J, Van der Vleuten C 2010 A standard setting method with the best performing students as a point of reference: Practical and affordable. Med Teach 32 154 – 160

http://dx.doi.org/10.3109/01421590903196979

Downing S, Tekian A, Yudkowski R 2006 Procedures for establishing defensible absolute passing scores on performance examinations ion health professions education. Teach Learn Med 18: 50 – 57

http://dx.doi.org/10.1207/s15328015tlm1801_11


http://dx.doi.org/10.1046/j.1365-2923.2003.01495.x


http://dx.doi.org/10.3109/0142159X.2011.611192

Appendices
Declaration of Interest

The author has declared that there are no conflicts of interest.