RMJ Ramathibodi Medical Journal

Rama Med J | Special Article

Fixed Scale-Referenced Versus Minimal Passing Level-Referenced Scaling Methods for Grading: The Perspective for Practice

Permphan Dharmasaroja

Chakri Naruebodindra Medical Institute, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Samut Prakan, Thailand

Setting standards is an important aspect of a grading system, which should be fair, reasonable, practical, and transparent. Distinct types of assessments result in different score distributions and different pass grades, and substantial errors can occur during aggregation of scores from multiple types of assessment. Thus, converting scores to a common grade structure is therefore essential. Different methods for the final grading of a course can cause the potential of disparity in course grades. The present work provides illustrative examples of 2 fixed scale-referenced grading methods and the minimal passing level-referenced grading method to demonstrate the potential of discrepancy of different grading scales. Perspectives for practice are provided.

Keywords: Grading, Minimal passing level, Assessment

Rama Med J: doi:10.33165/rmj.2022.45.2.255947

Received: January 31, 2022 **Revised:** March 31, 2022 **Accepted:** June 13, 2022

Corresponding Author:
Permphan Dharmasaroja
Chakri Naruebodindra Medical Institute,
Faculty of Medicine Ramathibodi Hospital,
Mahidol University,
111 Moo 14 Suwannabhumi Canal Road, Bang Pla, Bang Phli, Samut Prakan 10540, Thailand.
Telephone: +66 2839 5161
E-mail: permphan.dha@mahidol.ac.th

Introduction

The final grades that students receive in intra-curricular courses define their academic success in a program. The course instructor assigns a letter or number grade to summarize all evaluations of a student's performance during the course, resulting in these grades. Setting standards is an important aspect of an evaluation strategy. The method of standard-setting should be fair, reasonable, practical, and transparent, and the passing criterion should be legitimate,[1] passing students who are truly competent and failing those who are not. It is commonly understood that there is no such thing as a "gold standard" approach to standard setting, and that different approaches give varied results.[2] The motivation for exploring the assignment of grades came from the practice on the grading method adapted at the Preclinical section in Ramathibodi Medical School, Faculty of Medicine Ramathibodi Hospital. The results of the practice highlighted the impact of different grade scaling on the number of students who passed or failed a course. This article is aimed to demonstrate the potential of discrepancy of different grading scales through illustrative examples.

Marks, Scores, and Grades

The term "*mark*" combines 2 notions that can be distinguished as *scores* and *grades*. A "score" is a raw performance data obtained from a test. A "grade" indicates the level of performance or achievement.[3] Distinct types of assessments result in different score distributions and may give different pass grades. The score can be translated to a grade in each particular case by using a conversion algorithm. As different methods of assessment produce distinct score distributions, substantial errors can occur during aggregation of scores from multiple types of assessment. Converting scores to a common grade structure is therefore essential.[4] However, choosing a conversion algorithm might be a difficult issue, especially if both the scoring and grading systems are non-interval. Typically, there is a wider range of multiple choice question (MCQ) scores than with some other kinds of assessment such as essay writing. Converting essay markings will generally pull in the tails of an MCQ score distribution.[3] In addition, MCQ scores would only represent an interval scale if all the questions were of exactly the same difficulty, or if differential marks were awarded to questions on the basis of their difficulty.[3]

Grade Aggregation Methods

Aggregation is the procedure of calculation to provide single descriptions of student performances from diverse assessments,[3] and it determines how grades in a category are combined. Several aggregation methods have been described such as natural weighting, mean of grades, weighted mean of grades, and median of grades. One of the most common used aggregation methods is 'weighted mean', also called 'weighted average letter grade'. By using this method which allows assessment tasks to be marked out of 100 (or whatever),[5] instructors convert the score on each assessment task into the numerical equivalent of a letter grade, using a variety of scale such as the 4.0 scale. After that, the instructor calculates a weighted average of all of the assignments and tests. This method of calculating grades of assessments within a course is similar to how universities calculate overall student grade point averages (GPA) by combining grades from various courses.

Considering the scenario shown in Table 1, assume a student received a 55 on the MCQ exam. Using the criteria established by the course, the student has earned the 2.0 point (or C grade) and the grade by item weighted is calculated as: $2 \times 50/100 = 1$. A similar calculation is applied to other assessments. There is a noteworthy issue arising in this scenario that is how a score is assigned as a 'grade point'. This circumstance requires a standard setting that should be fair, reasonable, practical, and transparent. A common solution is to use the cut-off score (or minimal passing level) that a student would be expected to achieve to pass the assessment, in combination with appropriate interval scales.

Table 1. Item Weight in 'Weighted Mean' Aggregation Method

Grade Point on 4.0 Scale (Grade Equivalent)	MCQ Exam		Lab Exam		Assignment	
	Score	Item Weight 50%	Score	Item Weight 20%	Score	Item Weight 30%
4.0 (A)	85 - 100	2	87 - 100	0.8	91 - 100	1.2
3.5 (B+)	77 - 84	1.75	78 - 86	0.7	83 - 90	1.05
3.0 (B)	69 - 76	1.5	69 - 77	0.6	75 - 82	0.9
2.5 (C+)	61 - 68	1.25	60 - 68	0.5	67 - 74	0.75
2.0 (C)	53 - 60	1	51 -59	0.4	59 - 66	0.6
1.5 (D+)	45 - 52	0.75	42 - 50	0.3	51 - 58	0.45
1.0 (D)	37 - 44	0.5	33 - 41	0.2	43 - 50	0.3
0 (F)	≤ 36	0	≤ 32	0	≤ 42	0

Abbreviation: MCQ, multiple choice question.

Minimal Pass Level

Standard setting is a fundamental process in defining competency and level of learning performance. However, there is yet no gold standard method of standard setting. Several methods of systematic judgement to set minimal passing levels (MPL) have been widely used in medical education. One commonly used method is the Angoff method in which a group of subject-matter experts estimates the percentage of borderline students predicted to correctly answer each question in an examination.[6] In this method, the experts judge how difficult each item is in an exam to determine the MPL. The Angoff method is costly, time consuming and relies on the assumption that the expert panel can accurately define the borderline student. The Angoff method is a test-centered, predefined criterion-referenced method. The modified Angoff method has been developed to allow panelists given information such as test result and other panelists' rating result to discuss the cut-off score.[6]

In 2010, the Cohen method has been developed and subsequently modified in 2011. This simple and affordable approach is referred to as a compromise method since it combines components of both absolute methods (based on students' performance) and relative methods (where the number of passing students is relative to the rest of the students taking the exam). The Cohen's MPL is calculated by taking the student in the 95th percentile (the student whose score is higher than 95% of the rest of the students taking the same exam) and finding 60% of their score:[7]

$$\text{Cohen MPL} = 0.60 \times P95$$

A potential weakness of the Cohen method is that the multiplier (.60) and the spectrum of the reference group (P95) were seen as being somewhat arbitrary. The modified Cohen method estimates these values more accurately based on historical data from previous exams to come to a more reliable decision regarding the MPL value. The modified version was established by changing the 95th percentile to 90th percentile, since students in the top 10% of the tests responded differently to exam difficulty than other students and the student performance was consistent over time holds for the exams. The multiplier was also changed from 0.60 to 0.65, based on the ratio of the cut score to the score of the student at the 90th percentile on exams that have been standard set using the modified Angoff method:[8]

$$\text{Modified Cohen MPL} = 0.65 \times P90$$

As can be seen from the above, there are several methods for determining the exam MPLs. There are 2 main categories of methods to determine the MPL: criterion-referenced methods and norm-referenced methods.[9] Criterion-referenced methods (also known as test-centered

standard setting) which include the Angoff method, are independent of the test results. On the other hand, a norm-referenced method establishes its MPL in the form of a pre-determined score of the test items that need to be correctly answered. The norm-referenced methods include the Cohen and modified Cohen methods. It is widely accepted that there is no such thing as a "gold standard" approach to standard setting.[2]

Score-to-Grade Conversion

The number of grade points used, the intervals between grades, and whether they are represented as letters, numbers or other descriptors, differ widely between medical schools. The range of scores is dictated by the nature of the assessment. It is recommended that all scores should be converted to grades (ie, standardized scores) before aggregation, and conversion algorithm should be set with regard to each individual assessment.[3]

One common practice is norm-based referencing of assessments. However, the score distribution varies with the number of factors. These factors include the difficulty level of the test, the intellect and diligence of the students, how well the students were aided while learning, and how the assessments were marked. All of these may vary from year to year. If the assessment varies more than the students' characteristics, norm referencing may be more valuable. If the students vary more than the assessment, then criterion-based referencing may be more suitable.[10] There are also compromise systems that aim to combine the benefits of both norm- and criterion-based referencing.

Normally, the lowest expected score can be set to the equivalent to the lowest grade, and the highest expected score expected from a fully prepared and capable student to achieve can be set as equivalent to the highest grade. Once the 3 fixed points (MPL, lowest and highest scores) have been established, then the intervening scores or grades can be created. Another approach to create the intervals of the scores is by using the MPL as a reference point to assign grades above and below the pass grade such as a scenario shown in Table 2. Assuming that the calculated MPL that is the cut-off score for the grade point '2' is 53.12 and a predetermined interval is 7.5% of the total score of the MCQ exam, then the score interval for the grade point '2' is between MPL to < MPL + 7.5% (ie, 53.12 - 60.61). The same method is applied to other score intervals above and below the MPL. One criticism of this method is the value of the predetermined interval that is arbitrary. Other intervals have also been applied, and other methods of interval setting can be used such as the Dewey method.[11]

Table 2. Criteria for Grading Using Percentages of the Total Assessment Score

Criteria	MCQ Exam		
	Score (Total = 100)	Grade Point on 4.0 Scale	Item Weight 50%
≥ MPL + 30%	83.12 - 100	4.0	2
MPL + 22.5% - MPL + 29.9%	75.62 - 83.11	3.5	1.75
MPL + 15% - MPL + 22.4%	68.12 - 75.61	3.0	1.5
MPL + 7.5% - MPL + 14.9%	60.62 - 68.11	2.5	1.25
MPL - < MPL + 7.5%	53.12 - 60.61	2.0	1
MPL - 7.5% - < MPL	45.62 - 53.11	1.5	0.75
MPL - 15% - MPL - 7.4%	38.12 - 45.61	1.0	0.5
< MPL - 15%	< 38.11	0	0

Abbreviations: MCQ, multiple choice question; MPL, minimal passing level.

Setting the Final Grades

Regarding the scenario in Table 1, suppose that a student obtained a weighted point of 1.5 on the MCQ exam, 0.7 on the lab exam, and 1.05 on the assignment task, resulting in a total weight point of 3.25, what the final grade of the course the student should receive. Should the intervals of the final grade of the course be the same as the intervals of the grade points used in each assessment of the course? What criteria should be used in setting the final grades? It is unusual for grade scales to be symmetric around the pass grade (or MPL). Frequently, there are more grades above the MPL than below it. A principle of continuity could be used,[3] which requires that even increments are employed between the lowest grade and the MPL, and separately between the MPL and the highest grade. However, grades may be concentrated in areas with a large density of student scores.

Illustrative Examples

Three scaling systems are created as examples for grading of a course (Table 3), which include the following:

Example 1: Fixed scale-referenced grading system with uneven interval. This system has equal intervals of 0.50, except at B to B+ and B+ to A where the intervals are 0.25.

Example 2: Fixed scale-referenced grading system with even interval. Under this scheme, each grade is separated by a 0.50 interval.

Example 3: Minimal passing level-referenced grading system with even interval. This system is based on the Cohen-based MPL of a course and has an equal interval of 7.5% of the overall scores of the course.

To illustrate the potential of disparity in course grades, a single dataset of scores was generated for MCQ exam, lab exam, and assignment task, with item weight of 50%, 20%, and 30%, respectively. Each dataset, which is not actual course data, is generated in Excel with respect to the following values:

- MCQ exam: size of dataset is 200, with the total score of 100, the mean score of 73 and the standard deviation of 11.

- Lab exam: size of dataset is 200, with the total score of 100, the mean score of 67 and the standard deviation of 12.

- Assignment task: size of dataset is 200, with the total score of 100, the mean score of 89 and the standard deviation of 4.

All values used to generate the dataset was based on the actual values of the real course (RAID204: Clinical Anatomy course). The scores of each dataset were then converted to grade points using the weighted-mean aggregation method, with the MPLs (equivalent to the grade point '2.0') calculated by the standard Cohen's method. Grade points for each assessment were then obtained using the criteria in Table 2. The total grade point of a student was obtained by summating of grade-point values from each assessment task (ie, MCQ + lab + assignment). Therefore, the MPL of the course was summating of the MPLs of each assessment, which is equal to 2.25 (from a 4.0 scale). When applied to the scaling system for course grading, the number of students who receives each grade is shown in Table 4.

Two major issues should be noted from these illustrative examples. First is the number of students who failed the course. It is observed that MPL-referenced grading can detect students who should fail the course (grade below C) more than the other 2 fixed scaled-referenced methods. All of the failed students have low raw scores (44% - 54%) in both MCQ and lab exam, which are the knowledge domain of the course. Moreover, this observation is similar to what was discovered in RAID204, the real preclinical course in the academic year 2021. Thus, it is plausible that fixed scale-referenced grading methods produce a false positive error (passing a student who should have failed).

Second is the number of students above the pass grade 'C'. There is no students who get 'A' in the fixed scale-referenced method with 'even' interval. This may be a reason that one prefers the fixed scale-referenced method with 'uneven' intervals in order to distinguish students with high performance. However, choosing this method may raise the question of whether it is fair, reasonable, and

transparent, comparing with the MPL-referenced grading method with even interval.

Compared with the Cohen method, applying the modified Cohen method to the same dataset shows a similar pattern of findings, although the MPL is changed from 2.25 to 2.19 (Table 5). However, the modified Cohen method seems to reduce the number of students who should get higher grades by pulling down toward the pass grade. A potential of the Cohen method to inflate the risk of a false positive error (getting a higher grade than it should be), compared with the modified Cohen method, remains to be explored. A previous study suggests that the score of the student at the 95th percentile in the Cohen method may not be the most appropriate reference point, depending on the shape of the distribution of students' scores. In contrast, when using the Cohen method the performance of the 90th-percentile student is consistent over time.[8]

Table 3. Scaling System for Grading

Grade	Fixed Scale-Referenced With Uneven Interval	Fixed Scale-Referenced With Even Interval	MPL-Referenced With Even Interval
A	> 3.5	4.00	≥ MPL + 30%
B+	3.25 - 3.49	3.50 - 3.99	MPL + 22.5% - MPL + 29.9%
B	3.00 - 3.24	3.00 - 3.49	MPL + 15% - MPL + 22.4%
C+	2.50 - 2.99	2.50 - 2.99	MPL + 7.5% - MPL + 14.9%
C	2.00 - 2.49	2.00 - 2.49	MPL - < MPL + 7.5%
D+	1.50 - 1.99	1.50 - 1.99	MPL - 7.5% - < MPL
D	1.00 - 1.49	1.00 - 1.49	MPL - 15% - MPL - 7.4%
F	< 1.00	< 1.00	< MPL - 15%

Abbreviation: MPL, minimal passing level.

Table 4. The Number of Students in Different Grading Scales With MPL Based on the Standard Cohen Method

Grade	Fixed Scale-Referenced With Uneven Intervals		Fixed Scale-Referenced With Even Intervals		MPL-Referenced With Even Intervals	
	Scale	No. of Students	Scale	No. of Students	Scale	No. of Students
A	> 3.5	56	4.00	0	> 3.45	64
B+	3.25 - 3.49	44	3.50 - 3.99	56	3.15 - 3.44	59
B	3.00 - 3.24	50	3.00 - 3.49	94	2.85 - 3.14	44
C+	2.50 - 2.99	43	2.50 - 2.99	43	2.55 - 2.84	25
C	2.00 - 2.49	6	2.00 - 2.49	6	2.25 - 2.54	4
D+	1.50 - 1.99	1	1.50 - 1.99	1	1.95 - 2.24	3
D	1.00 - 1.49	0	1.00 - 1.49	0	1.65 - 1.94	1
F	< 1.00	0	< 1.00	0	< 1.65	0

Abbreviations: MPL, minimal passing level.

Table 5. The Number of Students in Different Grading Scales With MPL Based on the Modified Cohen Method

Grade	Fixed Scale-Referenced With Uneven Intervals		Fixed Scale-Referenced With Even Intervals		MPL-Referenced With Even Intervals	
	Scale	No. of Students	Scale	No. of Students	Scale	No. of Students
A	> 3.5	31	4.00	0	> 3.39	43
B+	3.25 - 3.49	40	3.50 - 3.99	31	3.09 - 3.38	47
B	3.00 - 3.24	39	3.00 - 3.49	79	2.79 - 3.08	55
C+	2.50 - 2.99	71	2.50 - 2.99	71	2.49 - 2.78	36
C	2.00 - 2.49	16	2.00 - 2.49	16	2.19 - 2.48	14
D+	1.50 - 1.99	3	1.50 - 1.99	3	1.89 - 2.18	3
D	1.00 - 1.49	0	1.00 - 1.49	0	1.59 - 1.88	2
F	< 1.00	0	< 1.00	0	< 1.59	0

Abbreviations: MPL, minimal passing level.

Perspectives for Practice

It is widely agreed that standard-setting for grading should be fair, reasonable, practical, and transparent. Evidence-based approach should also be taken into account. The following perspectives apply where 'weighted mean' aggregation of marks across multiple types of assessments is undertaken, meaning that scores from all types of assessment are converted to a common grade scale (standardized score) before aggregation and later decision for the final grading of a course.

1) Assuming that an MPL is derived from a well-established algorithm either test-centered or performance-centered, the MPL-referenced grading system with even intervals may be more justified and explainable than the fixed scale-referenced methods.

2) Whether to use the modified Cohen method or the standard Cohen method may require testing using local data, especially to address the risk of false positive and false negative error rates. However, the modified method has been shown to reduce the subjectivity of the standard method.[8]

3) The inappropriate setting of the passing scale in the fixed scale-referenced grading method may raise the probability of a false positive error (passing a student who should have failed). To justify this inaccuracy, the MPL-referenced grading system with even intervals should be adopted, or used as a supplementary method.

Furthermore, research to evaluate how different grading scales affect subsequent false positives and false negatives is needed.

Conclusions

The nature of the grading scale may differ from one institution to the others. It is unknown whether there is a single best grade scale. However, a 'one-size-fits-all' grading scale should be avoided because individual courses are all different, either on the difficulty levels of the tests or the nature of the assessment. Whatever the grading system is used, medical schools should ensure that students who are judged 'just adequate' are in fact safe doctors when they start their careers.

References

1. Norcini JJ. Setting standards on educational tests. *Med Educ.* 2003;37(5):464-469. doi:10.1046/j.1365-2923.2003.01495.x
2. George S, Haque MS, Oyebode F. Standard setting: comparison of two methods. *BMC Med Educ.* 2006;6:46. doi:10.1186/1472-6920-6-46
3. McLachlan JC, Whiten SC. Marks, scores and grades: scaling and aggregating student assessment outcomes. *Med Educ.* 2000;34(10):788-797. doi:10.1046/j.1365-2923.2000.00664.x
4. Vella F. A handbook for teachers in universities and colleges: a guide to improving teaching methods: by D Newble and R Cannon. pp 159. St Martin's Press, New York. 1989 ISBN 0-312-03196-3. *Biochemical Education.* 1990; 18(4):211. doi:10.1016/0307-4412(90)90143-C
5. Monash University. Learning and Teaching: Teach HQ: Weighting aggregation, 2022. Accessed January 15, 2022. https://www.monash.edu/learning-teaching/teachhq/moodle/gradebook/how-to/weighting-aggregation
6. Angoff WH. Scales, Norms, and Equivalent Scores. In: Thorndike RL, ed. *Educational Measurement.* 2nd ed. American Council on Education; 1971: 508-600.
7. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. *Med Teach.* 2010;32(2):154-160. doi:10.3109/01421590903196979
8. Taylor CA. Development of a modified Cohen method of standard setting. *Med Teach.* 2011;33(12):e678-e682. doi:10.3109/0142159X.2011.611192
9. Ben-David MF. AMEE Guide No. 18: standard setting in student assessment. *Med Teach.* 2000; 22(2):120-130. doi:10.1080/01421590078526
10. Lowry S. Assessment of students. *BMJ.* 1993;306(6869):51-54. doi:10.1136/bmj.306.6869.51
11. Arunyingmongkol S, Laisnitsarekul B. SB-8: the eight-category grading program. *Chula Med J.* 1998;42(7):587-597.

RMJ Ramathibodi Medical Journal

Rama Med J | Special Article

วิธีจัดช่วงระดับในการตัดเกรดที่อ้างอิงมาตราส่วนคงที่เทียบกับอ้างอิงระดับผ่านขั้นต่ำ: มุมมองสำหรับการปฏิบัติ

เพิ่มพันธุ์ ธรรมสโรช

สถาบันการแพทย์จักรีนฤบดินทร์ คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล สมุทรปราการ ประเทศไทย

การกำหนดมาตรฐานถือเป็นส่วนสำคัญของระบบการตัดเกรด ซึ่งควรมีความเป็นธรรม สมเหตุสมผล นำไปใช้ได้จริง และโปร่งใส การประเมินด้วยรูปแบบที่แตกต่างกันส่งผลให้มีการกระจายของคะแนนที่แตกต่างกัน และได้คะแนนผ่านเกณฑ์ที่แตกต่างกัน ความคาดเคลื่อนที่สำคัญอาจเกิดขึ้นในการรวมคะแนนจากการประเมินหลายประเภท ดังนั้น การแปลงคะแนนเป็นโครงสร้างเกรดรูปแบบเดียวกันจึงเป็นสิ่งจำเป็น วิธีการตัดเกรดขั้นสุดท้ายที่แตกต่างกันของรายวิชาอาจทำให้เกิดความเหลื่อมล้ำของเกรดในรายวิชานั้นได้ รายงานนี้ได้แสดงตัวอย่างของวิธีการตัดเกรดที่อ้างอิงมาตราส่วนคงที่ 2 วิธี และวิธีการตัดเกรดที่อ้างอิงระดับผ่านขั้นต่ำ เพื่อแสดงให้เห็นถึงความเป็นไปได้ที่อาจเกิดความไม่สอดคล้องของการตัดเกรดที่มีช่วงระดับแตกต่างกัน อีกทั้งรายงานนี้ได้ให้มุมมองในการปฏิบัติด้วย

คำสำคัญ: การตัดเกรด ระดับผ่านขั้นต่ำ การประเมิน

Rama Med J: doi:10.33165/rmj.2022.45.2.255947

Received: January 31, 2022 **Revised:** March 31, 2022 **Accepted:** June 13, 2022

Corresponding Author:
เพิ่มพันธุ์ ธรรมสโรช
สถาบันการแพทย์จักรีนฤบดินทร์
คณะแพทยศาสตร์
โรงพยาบาลรามาธิบดี
มหาวิทยาลัยมหิดล
111 หมู่ 14 ถนนเลียบคลอง
ส่งน้ำสุวรรณภูมิ
ตำบลบางปลา อำเภอบางพลี
สมุทรปราการ 10540 ประเทศไทย
โทรศัพท์ +66 2839 5161
อีเมล permphan.dha@mahidol.ac.th