ASIAN JOURNAL OF THE SCHOLARSHIP OF TEACHING AND LEARNING VOLUME 6 ■ NUMBER 1 ■ MAY 2016

ORIGINAL ARTICLE

Exploring the Aetiology of Grade Moderation: Is there a place for prophylaxis?

HO Han Kiat[1]

[1] Department of Pharmacy, National University of Singapore

Address for Correspondence: Dr Ho Han Kiat, Dept of Pharmacy, Faculty of Science, National University of Singapore, 18 Science Drive 4, Singapore 117543

Email: phahohk@nus.edu.sg

Recommended citation:

Ho H. K. (2016) Exploring the aetiology of grade moderation: Is there a place for prophylaxis? *Asian Journal of the Scholarship of Teaching and Learning, 6*(1), 27-46.
https://doi.org/10.24112/ajsotl.63325

Exploring the aetiology of grade moderation: Is there a place for prophylaxis?

ABSTRACT

Grade moderation is a deliberate exercise conducted in most institutions and at various levels of the student evaluation process that seeks to facilitate quality assurance of the educational program at large. The author takes a closer look at the practice of grade moderation that operates specifically at the modular or program level where the aim is to differentiate student abilities by banding them into different performance levels, in order to conform to an expected, consistent and equitable standard. This approach is more commonly known as curve fitting, by conforming to an expected reference norm for performance distribution in a relatively large population. The purpose of this paper is to critically consider the premise for such moderation, the objectives and the social repercussions that accompany its implementation as an unintended outcome. Central to the discussion are the notions that: (1) the aetiologies which necessitate grade moderation as a response to achieve the reference norm, are largely preventable in nature; (2) the assessor contributes fundamentally to the aetiologies, and therefore should be the targeted party for such prevention. Ultimately, the author advocates the personal responsibility of the assessor in monitoring student performance formatively, and the need to exercise careful considerations in the crafting and grading of assessments. This will help to minimise the use of institutional grade moderation as a necessary evil, which should otherwise be preserved only as a contingency measure when prophylaxis fails.

INTRODUCTION

Why grade moderation?

Grade is an integral part of assessment. It informs both the assessor and the student about the performance on the assessment task. In so doing, grade achieves two principal goals of assessment: (1) development of talents and (2) selection of talents (Guskey, 2011). Such goals are particularly relevant in small nation states like Singapore where human capital is the primary resource and a robust mechanism to develop and to select the best is essential for the economy. For this reason, quality grading must be upheld to fulfill both objectives. Inadvertently, grade moderation becomes an instrument employed during the process of administering assessments, occurring after grading or marking has been done. Institutionally, moderation facilitates quality assurance of the overall educational program (Sadler, 2013). Grade moderation is a process of calibrating grades both individually (i.e. student performance) and collectively (class performance), with the intention of achieving consistency and fairness according to institutional standards. Such practices safeguard against variability in performance standards arising not from the natural distribution of students' abilities (Chapman & Hills, 1916; Wedell, Parducci, & Roman, 1989), but from the subjective grading tendencies of individual assessors: for instance, whether assessors are too strict, too lenient, narrow distribution of grades, or simply erratic and unpredictable.

At the organisational level, grade moderation seeks to deliver a comparable standard both horizontally across modules, and vertically within the same module from year to year, reflected through a calibrated and normalised class performance. Such practice is true when a norm-referenced assessment context is adopted as the principle for grading. In the context of degree programs at large where students have the liberty to pick from a number of elective modules as they progress towards graduation, moderation of grades ensures equity in terms of the perceived level of difficulty between modules. This acts as a countermeasure against grade inflation introduced by the inconsistent or idiosyncratic grading patterns of specific assessors (Kulick & Wright, 2008). Furthermore, it could increase the likelihood that students would more likely choose a module based on interest and relevance, rather than the lure of the ease of achieving good grades in a particular module. At the societal level, such norm-referenced grade moderation facilitates the clear articulation of the performance levels of students. By assuming an expected distribution of the class into various levels of performance (for example 25% will achieve A-grade, 50% B grade and 25% C-grade), it helps to delineate the different achievement standards based on relative standing, rank profiles the cohort and supports the selection process for future employment of the graduates. In the context of

Singapore where the bulk of the graduates and job seekers come from a very small number of local institutions, grades becomes a common currency for comparing and selecting talents. Especially in cases where there are limited spaces and opportunities, ranking students based on academic performance is a useful tool to guide selection (McAllister, 1983). Hence, a quantifiable and reliable yardstick that is supported by a consistent and principled grade moderation should be sustained, so that sound decisions can continue to be made in the process of hiring.

Forms of grade moderation

The practice of grade moderation transcends various levels of the grading administration. At the ground level, there is an intuitive process of self-moderation, where the assessor repeats the marking in order to first check for accuracy of the answers, and then proceed to adjust the assigned grades as an attempt to correct any drifting tendencies over the course of grading several scripts (Lunz, Stahl, & James, 1989). This is particularly significant in large classes where one assessor can grade several hundred scripts or more within a short period of time. The grading of successive scripts may alter the assessor's interpretation of the acceptability of the answers as more variations in responses are observed. Therefore, the process of self-moderation offers an internal check and balance to ensure consistency in grading. Reciprocally, the intentional monitoring of student responses to the assessment provides feedback to the assessor about where learning challenges might be and how the articulation of the assessment problem might have perturbed their responses (or learning). Hence, this feedback practice generates as additional benefit through reinforcing the teaching-learning loop, consistent with the fundamental levels (i.e. first and second level) of training described by the Kirkpatrick learning model (Alliger & Janak, 1989). According to this model, learning is tiered and evaluated at the levels of the reaction of the student; the extent of knowledge and skills acquired; the application of the capabilities; and finally the outcome of the application on the business or environment.

Besides self-moderation, this internal grading audit can involve two or more assessors. Moderation arises when a second assessor re-grades the same response from the students, a model also known as consensus moderation (Linn, Burton, DeStefano, & Hanson, 1996). This provides a second opinion which either reinforces or challenges the verdict. The inputs of the additional assessor help to mitigate any subjective or even peculiar grading behavior of one assessor by spotlighting cases where dissonance between the graders' opinions are great or potentially contentious (Hunter & Docherty, 2011). After exercising this rigour, different treatments of the resulting grades can be applied. For instance, a grade can be derived as an average of the grades provided by the

two assessors; or if one of the assessors is the primary examiner, the second assessor can simply serve to flag out issues where a discrepancy in opinion is alarming, or the second assessor can contribute to a smaller percentage of the overall grade.

At a higher level, the overall class performance can be moderated. Such moderation considers the aggregate results of the class, and defines certain cut-off marks to distinguish between performance standards. Generally speaking, such efforts are directed to normalise the distribution performance standards across a population of students, often applied to educational institutions that subscribe to the practice of a norm-referenced assessment system. This process is also frequently referred to as "grading on a curve" (Wall, 1987). At the institutional or program level, this effort can be further adjusted to ensure comparability of course grades (Sadler, 2013). Yet, grading on a curve is a double-edged sword. This paper discusses the problems associated with the grade curve, and proposes a number of ways to overcome them.

Grading on a curve: Symptomatic relief that conjures a deeper problem?

Both the intent and the practice of grading on a curve can have tangible benefits for the institution and society, such as to aid in the selection of talents and the prevention of grade inflation. However, when such a curve fitting exercise results in a drastic and/or disproportional moderation of individual grades, it can generate a different set of adverse effects with unintended and negative impact on students. More critically, this paper seeds the notion that curve fitting merely provides symptomatic relief from issues of grade inflation, in lieu of an antidote targeted at the root of the problem. The primary aetiology may rest with the poor practices of some assessors. Therefore, addressing these deeper problems should be prioritised as a mitigating measure.

To contextualise this for the ease of discussion, I am going to use the model practiced in NUS, where a recommended percentage of students will be banded into specified categories of performance levels. I will discuss this perennial issue from a few angles; first, I will examine the social and ethical challenges associated with grading on a curve as the "side effects" of this practice (in Section 2 "Problems with Grade Moderation"); secondly, the paper will take a step back to consider the causes (i.e. aetiology) for a deviation from reference norms which drive the curve fitting exercise (in Section 3 "Aetiology for a Non-Gaussian Distribution of Class Performance"); and finally, I will consider the instrumental role the assessor plays in the whole paradigm and how an awareness of their role and responsibility can evoke a conscious attempt to instigate changes at the preventive level (in Section 4 "A Prophylactic Alternative to Curve Fitting").

PROBLEMS WITH GRADE MODERATION

In the process of moderating grades to fit the overall class performance into a profile prescribed by the institution, a number of unintended problems can arise. The overarching concern is whether this process may mask the authentic expression of academic achievements, and compromise the statutes of grade integrity. Sadler (2009) has proposed that maintaining grade integrity principally involves the following: (i) work should be graded strictly according to its quality without their responses compared with others in the group; (ii) grades should have comparable value across courses and institutions (p. 809). I will highlight four key issues and make some speculations on their repercussions.

A disproportional moderation of individual student's grades

Grade moderation, which entails fitting the class performance to certain historical profiles, can inadvertently lead to a disproportional moderation of individual student's grades. Let us take an arbitrary scenario where an ideal student population is expected to exhibit a reference distribution of grades with 25% of students achieving an A-grade, 50% a B-grade and the remaining a C-grade. We now assume that for one particular module, this distribution is achieved, but due to an "easier-than-usual" assessment task, the entire distribution shifts to the right such that 40% of the students achieve A-grade, 50% achieve B-grade, and the remaining a C-grade. To fit this reality back into the ideal profile, the cut-off score to achieve each grade category is now increased by a uniform margin. This lateral shift has now transformed the class performance back to the ideal state without perturbing the relative scores of each student. Arguably, this moderation has achieved its purpose of differentiating student performance, and conforms to an acceptable distribution by the institutional standards. Yet, the impact on each student is different. A student at the 60-70th percentile of the class experiences an unfortunate moderation downwards (receiving a B-grade for an otherwise A-grade performance); a student in the middle of the spectrum will experience no change; whereas a student at the lower end of the B-graders will be hit by an unexpected grade drop. Clearly, the mechanics of such moderation has adversely affected a pocket of students, even though the adjustment appears to be uniform and involves simply a lateral shift of the entire curve (Figure 1).

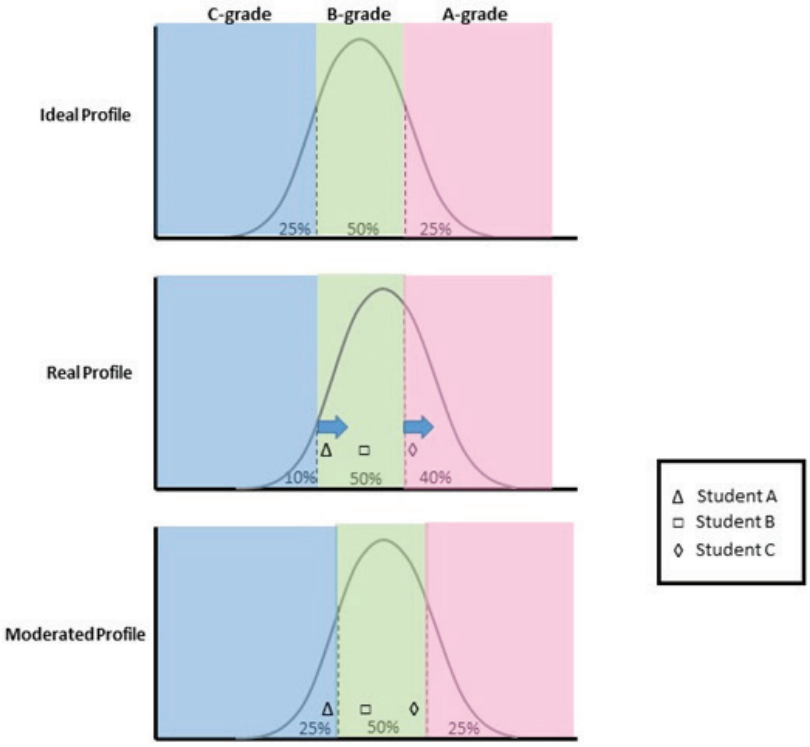

Figure 1. Effects of moderating class performance in a lateral left shift scenario.

Let us take another instance of a module where the assessment task has been unequivocally challenging. In this case, the students' grades are skewed towards the lower end of the spectrum with 10% of the students scoring an A-grade, 40% a B-grade and 50% a C-grade. Here, the performance profile has deviated from the reference distribution, and would necessitate a drastic measure of moderation at the lower end in order to re-establish the default profile. As a result, more students at the low end are bumped up to a higher grade through a recalibration exercise. While "re-shaping" the skewed profile back to normal, moderation has favored the weaker students and has blunted the original dissonance between the ends of the spectrum of student performance (Figure 2). More importantly, students who are right in the middle of the performance curve will not experience any change in grade. Though the mechanics of moderation is now different from the first example, the disproportional impact on specific subpopulations persists as an underlying inequality. Interestingly in this case, there appear to be some sweet spots of grades (e.g. for Students A and C) that will benefit some students but not others.

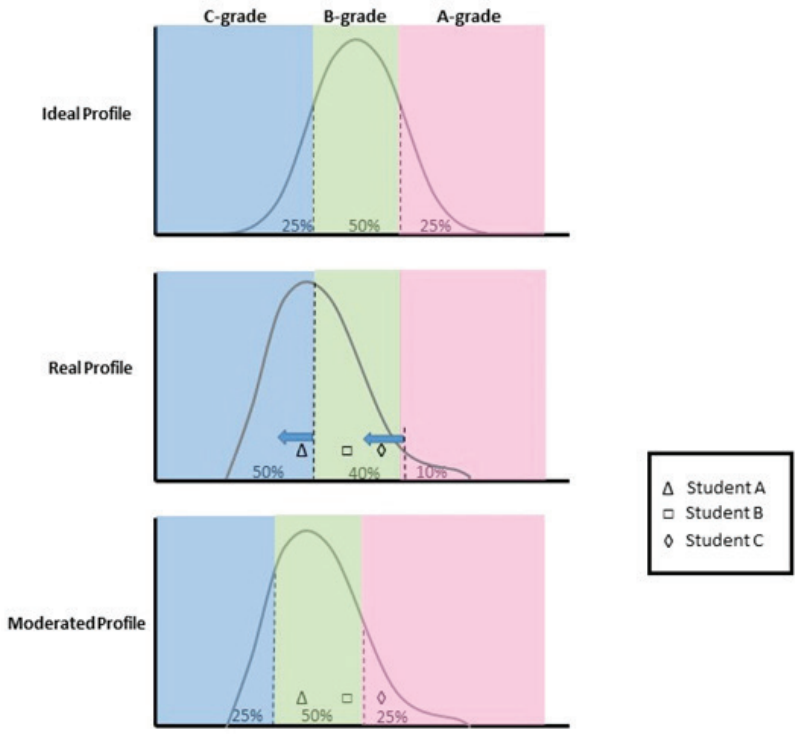

Figure 2. Effects of moderating class performance in a skewed right scenario.

An inexact summative feedback of true abilities

One of the key roles of grading is to provide a feedback channel for the students at the conclusion of the summative assessment, to inform them of their results. Therefore, moderating class performance at this stage can mask or even perturb the actual performance of individuals, and consequently distracts individuals from the knowledge of their academic performance which the adjusted grade has failed to inform. Let us consider two hypothetically extreme circumstances: In the first, Student A barely fulfills the criteria for an A-grade in a module. Based on the learning outcomes communicated, this grade signifies that the student has achieved the highest level of proficiency for the module. However, in a year where a large number of students receive an A-grade, the overall class performance may appear to have been skewed towards to the higher grades away from the norm. In order to achieve a normalised distribution of performance, the threshold for an A-grade is now elevated by a few percentage points. Student A, a borderline A-grader, unfortunately, has his/her grade adjusted downwards. When the feedback is delivered to Student A, he/she gets a rude

shock, namely that her performance is below par. Therefore, this apparent grade has misinformed the student of his/her performance. If it remains uncorrected, it can breed self-doubt regarding the student's mastery of content and a poorer outlook for subsequent performance in related modules.

On the flip side, let us consider a Student B, who is a weak student with poor understanding of the fundamental concepts that are needed to fulfill the learning outcomes of the subject. However, in one year where most students do poorly, Student B, together with several of the classmates, may get his/her grade inflated for the better, due to the moderation exercise where the threshold for performance is reduced to "shift the curve". While Student B would likely be surprised and thrilled with the results obtained, this feedback would give him/her a false sense of security, spurring him/her wrongly to take on more challenging courses subsequently that he/she may not be able to cope with.

Advocating an overly competitive culture

A norm-referenced grading characterised by grade moderation of class performance at the modular and program level, operates through scaling and adjusting thresholds in order to derive a desired distribution profile (e.g. Gaussian distribution). This distribution is necessary if the performance profile is ultimately used as a tool for ranking and selection of students into career tracks (Guskey, 2011). In other words, the decision to moderate, and the mode of moderation, depends on the performance of class relative to one another, rather than the absolute performance of individuals against a set of transparent criteria. The grades that the students received eventually will not have a direct bearing on how well they have learned (Guskey, 2011). Once again, we consider two extreme circumstances where the class is either uniformly good, or uniformly poor (i.e. narrow distribution with either a high or low mean). When we have an unusually gifted class where every student achieves excellence, such grade moderation forces some A-graders to take on lower grades. This creates a notion that A-grades are akin to limited collection trophies ("only one winner"), where individuals would need to "outshine the competition" in order to win them. Progressively, it can foster a culture of unhealthy competition, especially in a grade-driven society like Singapore, whereby students become more individualistic and less collaborative in their learning, and possessive of the knowledge they acquire (Krumboltz & Yeh, 1996; Shields & Bredemeier, 2010). Vice versa, in an unequivocally tough module where most students under-perform, grade moderation would inflate individual performance. Likewise with time, students may not be motivated to become truly proficient in the subject matter because an absolute measure of excellence is not necessary to win the competition. Using the analogy of the survival instinct among a herd of zebras chased by a ravaging lion, each zebra simply strives not to become the slowest runner of the herd. As long as they outrun the slowest zebra, they will

survive for at least another day. In this light, the danger is that those who are less intrinsically motivated by the content of the module may be more driven by the extrinsic influence of grades and simply aim not to be at the bottom of the pile. Overall, the relativity of grade determination creates an unhealthy competition among high-achieving students, and a spirit of mediocrity among low-achieving students.

Shifting the goalposts after the goal is scored

A final issue to consider is the idea that grade moderation at the module/program level is done as a post hoc exercise. Perceptually, it equates to shifting the goalposts after the goal is scored. In terms of implementation, the decision to moderate the class performance is often deliberated independently of individual student performance, but redresses abnormalities arising from the overall class performance. To some extent, this exercise also provides a platform for exceptional moderation measures to be taken for outliers, on a case-by-case basis. In other words, this level of moderation tends to supersede the performance standards that the assessors have already ascribed to individual students based on their actual responses to the assessment tasks and the relevant criteria. Take for instance a hypothetical case of a module where more than 90% of the students have achieved an A-grade in one particular year. This grade was justified by the assessors who felt that the class had indeed achieved the desired learning outcomes. Students' self-appraisal was also positive because they knew they could answer the questions well. However, final moderation led to many students receiving a poorer grade than anticipated. This outcome generated a sense of injustice and resentment against the educational institution. In an opposite scenario where students' grades were moderated upwards instead, while one may not expect grievances about the "improvement" in grades they were receiving, it would instead obfuscate the bases by which judgement were made about their performance. Taken together, this act of "grading on a curve" will inevitably shape the learning culture of the university (Erickson, 2011). It may have solved an administrative problem of grade inflation for the institution, but its execution has challenged the ethos of education in many ways (Biggs & Tang, 2011).

AETIOLOGY FOR A NON-GAUSSIAN DISTRIBUTION OF CLASS PERFORMANCE

From the multiple illustrations discussed so far, it is apparent that the crux of the problem is the existence of a discrepancy between the measured performance and the expected performance based on a reference norm. There are different manifestations of this discrepancy, which include but are not limited to, skewed-high, skewed-low, narrow, a double bell-shaped, and an erratic distribution.

Here, I would like to take an introspective view of these "symptoms" of discrepancies, and project some underlying "aetiologies" for their occurrences. But first of all, is a norm-based benchmarking for a population a realistic expectation in the case of student performance? Traditionally, this profile is based on historical and statistical consistency that a large number of students over a long period of time would lead to a grade reference distribution that can be used to predict future distribution (Al-Saleh, Ali, & Dahshal, 2010). Intuitively such an assumption would hold for as long as there is no change in the genetic makeup of the population, teaching methods, and that the subsequent population being compared with is sufficiently large in number. Undoubtedly, a small sample size would be a primary reason for the dissonance from expected trends, and therefore justifying it to be an exclusion criterion from curve fitting. However, we envisage that there are several other conditions that can perturb the normal trend, even within a large class size. This discussion focuses on such alternative scenarios.

Firstly, a class performance that is skewed-high can be due to the lack of rigor with the assessment questions (Linn et al., 1996). The assessment task may not be challenging enough to discriminate between good and truly exceptional performances. Let us use an example of an assessment comprising exclusively of multiple choice questions. Suppose that none of the test elements contained within these questions are highly differentiating of student abilities; in this case, it can be assumed that almost all could be answered by the majority of the students who have achieved the fundamental learning outcomes. This would lead to a clustering of the students' performance at the upper end of the spectrum. Such a situation could also be accentuated when there are insufficient test elements to help distinguish student abilities. This situation further highlights the danger of using just one mode of assessment against the criteria.

On the contrary, in a situation where students' performance is skewed-low, its aetiology could be nested on the premise of an unreasonably difficult assessment. Likewise, coupling this situation to insufficient test elements could similarly narrow the distribution around this mean. In addition, one may need to consider if the questions asked were fair, and within the context of intended learning outcomes. Failure of the majority of the class to achieve the required learning outcome should raise a cautionary flag as to whether there could be a gap between the desired learning outcomes and the instrument used to measure them (Bers, 2008).

Bunching around the mean refers to a class performance with a narrow distribution. This is another symptom of deviation from a reference distribution. While the overall profile conforms to normality, the real difference between the extreme performers may be too small for the performance standard to be

separated fairly and effectively. Presumably, this could lead to the system nitpicking on small differences in terms of the absolute performance, and transforming that into grade and ranking differences. Once again, this outcome misinforms the students of their true abilities. Yet bunching around the mean is a multi-factorial phenomenon that can arise as (1) a function of the assessor's grading pattern, or (2) the poor design of the assessment task. For example, in a free-response question (e.g. essay writing), an assessor may exhibit holistic grading and hence attribute general impression scores that clouds the differentiation between individual performances (Isenhour & Kramlich, 2008). As for poor question design and assessment format, narrow distribution of grades around the mean creeps in when there is only one obvious response students can provide. Furthermore, over-representation of group project work can be a source of results clustering because every member of the team will get the same score. The bigger the size of the group, and the bigger the apportioning of group work to the overall score, the greater the problem becomes.

Another significant deviation from normality is a double bell-shaped effect. This symptom underscores the absence of a homogeneous student population for the module concerned. It indicates that there could be two or more separate pools of students carrying innately diverse traits (or different backgrounds) that influence their performance (Yadin, 2013). For example, this can arise when there are prerequisites for a module and yet some students that do not meet the prerequisites are admitted into the class. This would then lead to a disparity in their background knowledge which shapes the learning curve differently. Furthermore, the specifics of the test elements can also contribute to diverse responses. For example, there could be differential interpretation of the assessment task by the students. Perhaps as a result of an ambiguous phrasing of the question, there are two possible responses, one leading them to a correct answer but the other causing them to pursue a completely different direction. Therefore, we end up with two distinct clusters of performance.

Finally, there are symptoms that are simply erratic and may not be suggestive of any specific underlying cause. Some of these could arise from inexperienced assessors who are not familiar with setting assessment problems and grading. In others, this could arise from some idiosyncratic behaviour of assessors. Independently, it could also be due to assessment tasks drawing multiple interpretations and responses, which the assessor has not anticipated at the point of setting the questions. Furthermore, the circumstances leading to a bi-modal distribution of performance as described above can further result in a more convoluted profile when there are more than two clusters of student backgrounds and abilities.

Taking all these symptoms and their likely aetiologies into consideration, one convergent point that emerges is the *pivotal role of the assessors*. Whether it is through shaping the framework of the assessment tasks, aligning the assessment problems with the learning outcomes, or administering the grading process, the assessor is instrumental to making a difference. Therefore, this motivates the main argument of this paper: whether we can obviate the subsequent use of curve fitting and avert its potential repercussions by tackling the root cause of the matter- i.e. the assessor's assessment knowledge and skills. In another word, is there a place for prophylaxis?

A PROPHYLACTIC ALTERNATIVE TO CURVE FITTING

To recapitulate the discussion, we alluded to two polarising realities: on one hand, the panoply of preconditions that trigger a deviation of class performance from a historical population norm (Section 3 of the paper), which then necessitates the use of curve fitting as a mitigating measure against the ensuing problems (in the "Introduction"); yet on the other hand, we recognise that any drastic moderation of grades to normalise class performance will inadvertently conjure a different set of concerns which can impact the social climate of the institution (in the second section of the paper). Therefore, without undermining the premise of grading on a curve for the purpose of rank-profiling and selection, we should as far as possible consider this as a backup plan while preventive measures can be prioritised. This leads to the foreground discussion of whether "prophylaxis" is available. We concluded Section 3 of this paper with the notion of the underpinning role of the assessor in the whole paradigm. Therefore, the cornerstone of a preventive measure is to begin with the assessor, both in terms of the design of the assessment tasks, as well as the grading practices. The rheostat that the assessor should constantly check against is the authenticity and stability of their assessment tasks towards achieving the intended learning outcomes. We will explore the specific ways an assessor can achieve this alignment in tandem with the various etiologies as delineated.

To deal with a skewed-high class performance, the broad-stroke strategy is to pitch the assessment activity at the right level. To a large extent, the crafting and selection of the questions requires some prior understanding of the student's profile and abilities. This is where a consistent review of formative assessment plays a major role (Frohbieter, Greenwald, Stecher, & Schwartz, 2011). Formative assessment can serve a dual purpose here: (i) It can help the students chart their progress and allows them to seek clarifications and corrections based on their mistakes in responding to the assessment tasks. When

the students perform well in such an assessment, it affirms their understanding and generates the confidence that they are on the right track towards attaining the learning outcomes for the subject (Yin et al., 2008). (ii) Furthermore, it can help the assessor gauge the level of understanding of the students, fine-tune or adjust the instructions and, in so doing, alter their learning journey prior to the summative assessment (Gurvitch & Lund, 2011).

Suffice to say, a well-designed formative assessment would be needed to achieve this positive effect. Ideally, it should be representative of the overall syllabus so that the performance indicators can reflect the true understanding of the students for the subject matter. An effective formative assessment should also be coupled to feedback so that learning gaps can be overcome and the students are better prepared for the final examinations (Chappuis, 2014). Reciprocal feedback from the students back to the instructor closes the teaching-learning loop for the instructor to understand the cause of the gap and appraise if future alteration of the content delivery or further clarification would be necessary. Overall, a thoughtful design of formative assessment with accompanying follow-up actions may prepare the students at large better so as to avert a situation of skewed performance. Therefore, this action is pre-emptive. In a way, this approach fulfils the the wider theoretical framework of the Kirkpatrick model which improves the training of students by negotiating the students' reactions and results, and eventually links them to a measure of subsequent job competency and performance level (Praslova, 2010).

Dealing with a skewed-low class performance can also benefit from a close study of formative assessments, however, with additional considerations. Where the majority of the class underperformed in the summative assessment, we need to critically consider if the results are due to poorly designed assessment tasks, or otherwise due to a lack of understanding of the subject matter by the students. In the former scenario, the questions may not have been interpreted in the manner the assessor has intended, in which case, the phrasing of the questions could be independently verified by another colleague. For the latter issue, we need to consider if the question is drawing prior knowledge that the students are not prepared for. In this case, the scope of the assessment and the congruence with the intended syllabus and learning outcomes should be re-evaluated. Like in the skewed-high case, the role of the assessor is essential and much can be done to align the assessment tasks to students' abilities.

Thirdly, the narrow distribution of the class performance around the mean complicates a decisive and fair moderation process. As discussed earlier, this trend could arise from an over-representation of group tasks, or it could be due to the "holistic grading" approach of the assessor. Each of these etiologies would require a different response. In modules where group tasks are instrumental to

the learning outcomes, different grouping arrangements can be applied for each subtask, so that we increase the diversity of the grouping and a "de-clustering" of the individual performance from the group performance at the closure of the module. As a word of caution, one option often taken to discriminate the performance of the students is to drastically increase the number of graded tasks to be completed within the assessment time frame. Presumably, this mechanism would widen the performance spread within a population. However, we need to be mindful whether such an approach is artificially stretching the performance distribution based on the ability of the students to complete their tasks within the specified time, rather than a measure of the difference in their learning abilities and competencies. Unless speed and accuracy of response are part of the learning objective of the module, this approach would not give a good representation of the students' abilities.

For narrow distribution arising from the peculiarity of the assessors, a set of well-defined and objective grading matrices could be instituted. This could include using more discrete grading components to prevent the situation of holistic grading by the assessor. The use of criterion-based assessment (or standards model) provides a tool to help distinguish individual attributes during grading (Biggs, 1995). In its practice, performance indicators are itemised into well-defined rubrics to guide the assessor to objectify their grading categorically. This introduces greater clarity whether a student fulfils a certain performance standard to warrant a specific grade. Such a grading approach will also help the assessor rationalise the grades awarded and provide useful qualitative feedback for the students.

Another frequently occurring phenomenon of a non-classical class performance is a profile with multiple bell-shaped distribution. Borrowing the concepts in epidemiology and genetics, such distribution suggests the existence of a mixed population carrying specific traits that directly impact the outcome being measured (or phenotype). Earlier, we have identified non-homogeneous background knowledge (or discrepancy in meeting pre-requisites for the module) as a source of difference. Another source of variability may arise from different linguistic abilities of the students. The earlier problem can be mitigated with more effective and accurate communication of the module description and assessment modes before the commencement of the learning journey. For an elective module where the enrolled students can come with different background knowledge, a pre-module self-analysis can be administered. Clear description of the module content as well as a sample of the assessment tasks can be included to help students considering the modules to evaluate their own suitability. Once the module has commenced, the instructor should continue to monitor the abilities of the students through in-class feedback, so that any divergence in abilities could be identified as early

as practicable. Linguistic challenges faced by some students are an inherent problem for any world-class university attracting a large number of non-native students who may not be proficient in the language of instruction. While lowering language standards is not an institutional solution to this problem, an assessor can take extra measures to ensure that the tasks prescribed for the students do not complicate the language used or introduce colloquial jargon that interfere with the non-natives' comprehension.

Finally, the grading idiosyncrasies of the examiners leading to a deviation from reference norm present a perplexing issue. As the cause for such erratic grading behavior can be highly varied, it is difficult to prescribe a one-size-fits-all recommendation for prevention. That said, a first step to alleviating this problem could begin with self-awareness. Recently, an exercise was conducted by the Department of Pharmacy at the National University of Singapore (NUS), whereby all staff were asked to grade four common and anonymised assessment reports of different performance standards. The grades provided by the staff were profiled and the relative stringency and leniency of the assessors were compared and revealed to the staff in confidence. This allowed the staff to recognise where they stood on the "grading stringency index", and also to identify grading outliers. Overall, it was the first time the staff received feedback of their grading pattern based on a relative scale as compared to a department average (*unpublished data*). This exercise has generated self-awareness as a starting point for future self-calibration and adjustment.

In addition, the institution needs to take on a larger role with regard to the proper induction, training and retraining of assessors and educators as a group. At NUS, the Centre for the Development of Teaching and Learning (CDTL) provides the foundational development of new teachers and additional resources for continuing professional development to help prepare educators for an ever-evolving teaching landscape. The tenets of good assessment tasks and the effective grading of student performance are part of what this unit can offer. Educators at NUS looking to enhance their assessment tasks in order to more effectively meet their course learning outcomes can approach the Centre for support, pedagogical expertise and opportunities for collaboration.

In summary, these aetiologies and their respective preventive measures are illustrated categorically in Table 1.

Table 1
Deviation of class performance from normality and preventive measures

Types of Deviation	Causes of Deviation	Preventive Measures
Skewed-high	Lack of rigour in assessment tasks	Increase rigour of assessment tasks through close monitoring of formative assessment performance, making adjustments to summative assessment where necessary.
Skewed-low	Highly challenging assessment tasks	Close monitoring of formative assessment performance, making adjustments to summative assessment where necessary.
	Misalignment between the intended learning outcomes and the assessments tasks	Obtain feedback from students with regards to their background; Carefully match learning outcomes with assessment, taking a third-party opinion for check and balance.
Narrow (bunching around the mean)	Holistic grading behaviour	Use criterion-based grading to clearly demarcate different performance levels.
	Over-representation of group tasks	Limit the percentage of group task in overall assessment; Use multiple grouping system to diversify the clustering of students and their results; Develop more rigorous group assessment methods that clearly defines grading criteria and policy about group or individual grades.
Double bell-shaped	Separate clusters of student background knowledge	Improve definition of the prerequisites and pre-module self-analysis
	Differential interpretation of the assessment tasks	Engage a third-party checker for the interpretation of assessment tasks.

Erratic	Multiple clusters of student background knowledge	Improve definition of the prerequisites and pre-module self-analysis.
	Idiosyncratic grading behaviour	Create opportunities for feedback and self-awareness of grading pattern.
	Inexperienced assessors	Provide proper induction, training and re-training for assessors; Encourage more conversations among colleagues especially those teaching in related courses.

CONCLUSION

In conclusion, this paper highlighted the premise for grade moderation in a norm-referenced assessment context, specifically targeting the course or institutional levels of moderation otherwise known as curve fitting. We considered how grade moderation is a measure to circumvent problems arising from grading abnormalities. However, we argued that many circumstances that warrant moderation are avoidable. To this end, we thrust the role of the assessor to the centre stage and deliberated on specific strategies to help increase the alignment of the class performance with the anticipated normality. This responsibility should not be undermined. Finally, to end off with an audacious and provoking thought, perhaps we should begin to ask ourselves as educators whether curve fitting exists because good assessors do not yet exist?

ACKNOWLEDGEMENTS

The author would like to thank colleagues including Assoc Prof Eric Chan, Assoc Prof Von Bing Yap and Dr. Hui Ting Chng for their critique, suggestions and proofreading of the manuscript.

ABOUT THE AUTHOR

Han Kiat HO is an Associate Professor at the Department of Pharmacy at the National University of Singapore. He is a toxicologist with research interests in the mechanistic understanding and management of drug-induced liver injury and other liver diseases. He is currently a fellow of the NUS Teaching Academy and is also the Deputy Head of the Department for educational matters.

REFERENCES

Al-Saleh, M. F., Ali, D., & Dahshal, L. (2010). Towards a reference curve for the grades of each course. *International Journal of Mathematical Education in Science & Technology, 41*(4), 547-555. http://dx.doi.org/10.1080/00207390903564645

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 42*(2), 331-342. http://dx.doi.org/10.1111/j.1744-6570.1989.tb00661.x

Bers, T. H. (2008). The role of institutional assessment in assessing student learning outcomes. *New Directions for Higher Education, 141*, 31-39. http://dx.doi.org/10.1002/he.291

Biggs, J. (1995). Assessing for learning: Some dimensions underlying new approaches to educational assessment. *Alberta Journal of Educational Research, 41*(1), 1-17.

Biggs, J., & Tang, C. (2011). Aligning assessment tasks with intended learning outcomes: Principles. *Teaching for Quality Learning at University* (4th Edition, pp. 191-223). England: Open University Press

Chapman, J. C., & Hills, M. E. (1916). A statistical study of the distribution of college grades. *The Pedagogical Seminary, 23*(2), 204-210. http://dx.doi.org/10.1080/08919402.1916.10534707

Chappuis, J. (2014). Thoughtful assessment with the learner in mind. *Educational Leadership, 71*(6), 20-26.

Erickson, J. A. (2011). How grading reform changed our school. *Educational Leadership, 69*(3), 66-70.

Frohbieter, G., Greenwald, E., Stecher, B., & Schwartz, H. (2011). *Knowing and Doing: What Teachers Learn from Formative Assessment and How They Use the Information*. CRESST Report 802 (pp. 57): National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Gurvitch, R., & Lund, J. (2011). The (missing) link between instruction and assessment. *Strategies: A Journal for Physical and Sport Educators, 25*(1), 32-34.

Guskey, T. R. (2011). Five obstacles to grading reform. *Educational Leadership, 69*(3), 16-21.

Hunter, K., & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment & Evaluation in Higher Education, 36*(1), 109-124. http://dx.doi.org/10.1080/02602930903215842

Isenhour, M., & Kramlich, G. (2008). Holistic grading: Are all mistakes created equal? *PRIMUS, 18*(5), 441-448. http://dx.doi.org/10.1080/10511970701624483

Krumboltz, J. D., & Yeh, C. J. (1996). Competitive grading sabotages good teaching. *Phi Beta Kappan, 78*(4), 324-326.

Kulick, G., & Wright, R. (2008). The impact of grading on the curve: A simulation analysis. *IJSoTL, 2*(2), 1-17. http://dx.doi.org/10.20429/ijsotl.2008.020205

Linn, R. L., Burton, E., DeStefano, L., & Hanson, M. (1996). Generalizability of New Standards Project 1993 pilot study tasks in mathematics. *Applied Measurement in Education, 9*(3), 201-214. http://dx.doi.org/10.1207/s15324818ame0903_1

Lunz, M. E., Stahl, J.A., & James, K. (1989). Content validity revisited: Transforming job analysis data into test specifications. *Evaluation and the Health Professions, 12*, 192-206. http://dx.doi.org/10.1177/016327878901200205

McAllister, E. (1983). Criterion-referenced versus norm-referenced measurement: Either or both? *Kappa Delta Pi Record, 19*(2), 58-60. http://dx.doi.org/10.1080/00228958.1983.10517719

Praslova, L. (2010). Adaptation of Kirkpatrick's four-level model of training criteria to assessment of learning outcomes and program evaluation in higher education. *Educational Assessment, Evaluation and Accountability, 22*(3), 215-225. http://dx.doi.org/10.1007/s11092-010-9098-7

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*(7), 807-826. http://dx.doi.org/10.1080/03075070802706553

Sadler, D. R. (2013). Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy & Practice, 20*(1), 5-19. http://dx.doi.org/10.1080/0969594x.2012.714742.

Shields, D. L., & Bredemeier, B. L. (2010). Competition: Was Kohn right? *Phi Delta Kappan, 91*(5), 62-67. http://dx.doi.org/10.1177/003172171009100516

Wall, C. R. (1987). Grading on the curve. *InCider, 5*(10), 83-85.

Wedell, D. H., Parducci, A., & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology, 102*(2), 233-248.

Yadin, A. (2013). Using unique assignments for reducing the bimodal grade distribution. *ACM Inroads, 4*(1), 38-42. http://dx.doi.org/10.1145/2432596.2432612

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Tomita, M., Furtak, E. M., Brandon, P. R., & Young, D. B. (2008). On the measurement and impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335-359. http://dx.doi.org/10.1080/08957340802347845 ■