AC 2010-504: GRADING TECHNIQUES FOR TUNING STUDENT AND FACULTY PERFORMANCE

Adrian Ieta, State University of New York, Oswego

Thomas Doyle, McMaster University

Rachid Manseur, SUNY-Oswego

Grading techniques for tuning student and faculty performance

New faculty are highly qualified in their own field, where they have accumulated some research experience and where they can bring fair amounts of enthusiasm. This article discusses grading techniques that help match student performance and instructor interest. Grading as a tool for evaluating student performance has been considered mainly from the student perspective. Anybody new to teaching rarely has proper training on grading techniques, which often are of least concern relative to teaching content. Nevertheless, grading as perceived by students may greatly impact their attitude towards the course and its instructor. It has been proven that students are very sensitive to grades and inaccurate evaluation of their perceived performance can also alter their future performance as well as their evaluation of teaching, which may adversely affect the instructor. Often, scaling of raw scores is used in grading engineering tests. There are no official standards on how this operation should be performed, hence the wide variation in the common procedures used. This work compares a few common-sense scaling procedures and shows how the final outcome may vary when determined from the same raw scores. Such grading variations significantly affect what the evaluation of the students' performance represents. This article offers recommendations on use of scaling methods so that the negative impacts of grading techniques and grade distortions can be minimized and lead to enhanced and efficient evaluation standards. Since grading and grading techniques are of general interest to instructors, this article may be of service to many instructors, especially to the new and relatively new faculty willing to review some of their own grading procedures.

Introduction

There are many aspects of academic life that new faculty have to deal with in a short period of time. Teaching, although demanding, is often a pleasant part. Marking papers and assigning grades are an essential component of the evaluation process, which may be more difficult than initially assumed. Moreover, student grades do bear weight on student evaluation of teaching (SET) scores [1], which may impact the future tenure and promotion of the faculty. Grading and student motivation for learning are related [2], although student motivation is not simply helped by high grades [3]. If the SET scores are not appropriate it is often very difficult to improve the scores without professional advice [4]. Some research shows that faculty can improve SET scores by giving higher grades [5-9].

In North America (but not only) the letter grade (LG) system is used in the student evaluation process. The LG system, developed at Harvard [10], has its advantage of carrying an easy intuitive meaning [11] and European Union now uses the European Credit Transfer and Accumulation System (ECTS) to convert European grades to LG categories [12]. However, this intuitive meaning evolves in time [3] and the interpretation of evaluation scores may vary according to the grading system used [13]. In this paper we aim at analyzing how grades are related to raw scores and how scaling can be used productively rather than harmfully in order to tune in student and faculty needs while following educational standards.

Score Scaling

There is no absolute meaning of a test raw score as percentages do not have to fit a particular grading scale used. Therefore, often scaling is needed in order to assign meaning to scores by means of grades. There are various methods for scaling. As there are no standards for scaling, one can use his/her method of choice arbitrarily. Sophisticated instructors may choose statistically intensive methods; however, most instructors use more mundane techniques. Some instructors known to the authors simply assign a letter grade to tests and aggregate LGs according to their frequency. More systematic approaches (yet common) to scaling and grading would commonly include some of the following methods [15]:

1. Straight scale (M1)

Let the raw score be denoted by "x" and the scaled score by "y". The scaled score is the raw score (identity function)

$y\% = x\%$ (1)

This method identifies the raw score as the "correct" score scale used by the instructor. If complex problems are used for tests, the likelihood of having low raw scores is higher than for a test assessing less complex, larger number of questions. Taking the raw score for granted may be unfair for evaluation. However, if the tests are well calibrated for such a grading approach, it may work well.

2. - Flat scale (a) (M2)

Let x_{max} be the maximum (percentage) raw score in a test set. Raw scores are translated according to (2)

$y = x + (100 - x_{max})$ (2)

Accordingly, $y_{max} = 100\%$. The method may be used for global adjustments to the grades. For instance, if nobody in the group was able to solve a particular problem the score associated with it may be added to the raw scores. Control over the average of a set of grades is limited by the value of the highest raw score, which may sometimes be a disadvantage.

3. Flat scale (b) (M3)

Raw scores are translated by a certain number but the highest scaled score and other scores can be larger than 100 %.

$y\% = x\% + b\%$ (3)

(b= an arbitrary percentage)

For instance, all but one raw score are within 35%-80% and there is one single score of 96%. According to the instructor's judgment, an x = 10% is added to all scores in order to obtain the numerical grade. Hence, 35 is scaled to 45 (= 35%+10%), 80 is scaled to 90 (= 80%+10%), and 98 is scaled to 106 (= 98%+10%). However, some instructors may not want to give scores larger than 100%. Nevertheless, instructors gain control on the class average grade.

4. Normalization to the highest score

Raw scores are normalized to the highest score.

$x = 100 * x / x_{max}$ (4)

If the highest raw score for a test is 96 this will be considered as 100%, hence a score of 66 will be scaled to 100*66/96 = 68.75). This method will boost high scores more than low ones. The method may be preferred when the scores are tight together but below the limit of an A, for instance. However, this scaling is normally disadvantageous to students with lower scores.

5. Linear Scale (M5)

$y = ax + b$ (5)

This method can be used to scale raw scores up to 100% or higher, if preferred. Using a coefficient "a" larger than 1 is conducive to more points added to the higher scores than to the lower scores, which is disadvantageous to students with poor performance. If a < 1 the lower scores are scaled by a larger amount of points. Let us assume that raw scores ranged from 30 to 80. Then for a=0.9 and b=20 we have $y(x_{max}=80) = 0.9*80+20= 92$ and $y(x_{min}=30) = 0.9*30+20= 47$. The minimum score was boosted by 17 points and the maximum by 12 points (please note that a=0.9<1 and 17>12). Similarly, for a=1.1 and b=10 we have $y(x_{max}=80) = 1.1*80+10= 98$ and $y(x_{min}=30) = 1.1*30+10= 43$. The minimum score was boosted by 13 points and the maximum by 18 points (please note that a=1.1>1 and 13<18). The "a" and "b" coefficients can be determined from equation (5) by chosing $y_M(x_M)$ and/or $y_m(x_m)$. Coefficient "a" is given by the ratio $(y_M-y_m)/(x_M-x_m)$. Setting a>1, it follows that (y_M-y_m) must be larger than (x_M-x_m). Coefficient "b" will be determined accordingly from equation (5).

As long as the scaled scores do not go over 100%, control of the average is somehow limited, which may be a disadvantage at times. The method is versatile, allowing for some control of the class average and for more boosting of lower scores than of higher ones, which is often needed.

6. Root function

Square root is applied to raw scores (%) and then multiplied by 10.

$y=10\sqrt{x}$ (6)

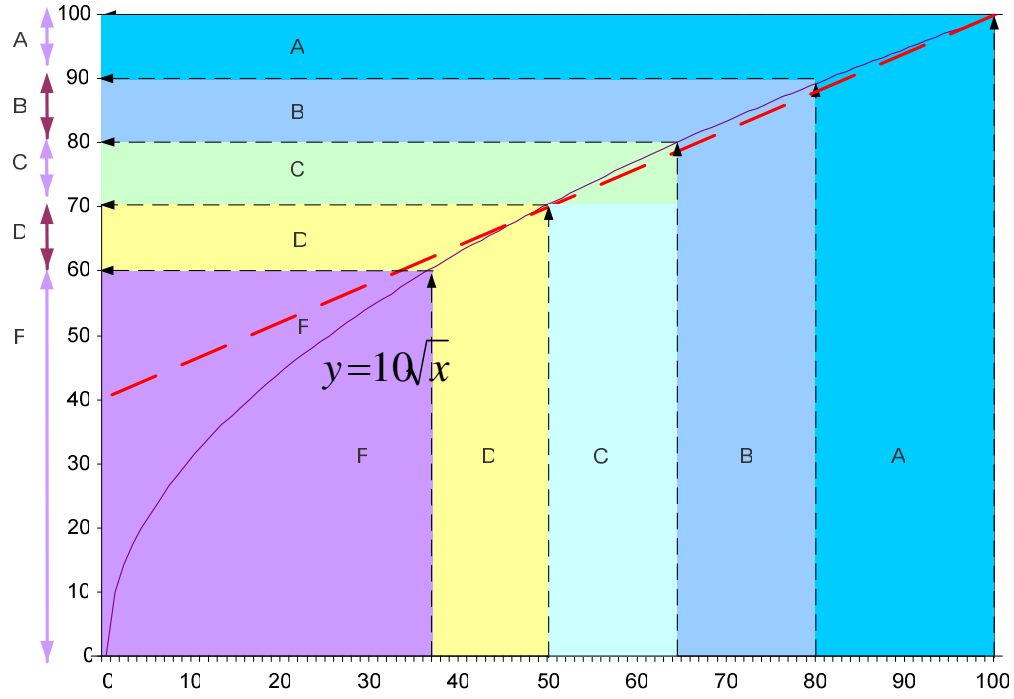


Figure 1. Root scaling method: conversion of raw scores (horizontal axis) using root scaling. The dotted red line shows that for most of the LG ranges the scaling can be very well approximated to a linear transformation.

For instance, a raw score of 81 (out of 100) will be scaled to $10\sqrt{81} = 90$. The method gives more boost to the lower scores than to the higher ones (see Fig. 1). It is apparent from Fig. 1 that for the 40 to 100 range the scaling is fairly close to a linear function (Method 5). However, coefficients "a" and "b" corresponding to Method 5 are fixed, which limits the versatility of this procedure. Some flexibility of the method can be achieved by repeating the scaling twice or more.
In addition to these popular methods, we promoted the following method, coined as Category-Based Linear Transformation Method.

7. The Category-Based Linear Transformation Method (CLT)

Let us assume that the A, B, C, D, and F LG categories are defined numerically by their border values. The CLT method performs linear transformations for each LG category so that:

$$y_i = \alpha_j * x_i + \beta_j \qquad (7)$$

where x_i and y_i are arbitrary raw scores and their scaled values correspond to each of the A, B, C, D, and F categories; α_j and β_j (j = 1, 2, …, 5) are linear transformation coefficients.

Let us assume that the instructor assesses the correspondence to each LG category of raw scores so that m_1, M_1 (minimum and maximum) correspond to the limits of the raw score of F LG; m_2, M_2 correspond to the limits of the raw score of D LG, etc.; also, on the projected scale we have F (m'_1, M'_1), D (m'_2, M'_2), C (m'_2, M'_2), etc. Then the coefficients α_j and β_j can be calculated as

$$\alpha_j = \frac{M'_j - m'_j}{M_j - m_j};\ \beta_j = \frac{m'_j M_j - m_j M'_j}{M_j - m_j} \quad (j = 1, 2, \ldots, 5) \qquad (8)$$

An example of scaling using this method is shown in Fig.2: the projected scale is defined by A (m_5= 90, M_5= 100), B (m_4= 80, M_4= 90), C (m_3= 70, M_3= 80), D (m_2= 60, M_2= 70) and the raw scale by A (m'_5= 80, M'_5= 100), B (m'_4= 60, M'_4= 80), C (m'_3= 40, M'_3= 60), D (m'_2= 30, M'_2= 40). The figure shows that for each LG we have a linear transformation of the raw score in the projected scale used by the instructor.

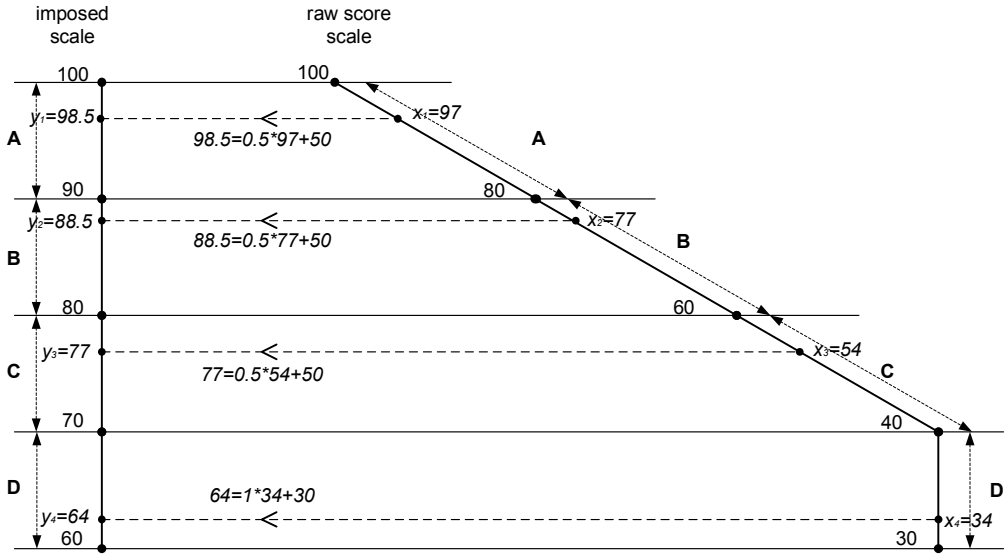


Figure 2. CLT scaling: raw scores (right side can be geometrically scaled to LG ranges corresponding to the CLT method (left side) [16]

The limits for LG categories can be chosen easily if scores cluster within certain values. However, if no specific clusters are noticed, the instructor should assign the LG limits according to his/her judgment and standards relative to the test, avoiding over-scaling. This method, tested by the authors within the last 5 years, offers the greatest flexibility from all the methods mentioned. It is more difficult to set up than the previous methods, but the efforts are rewarded as grades can be tuned to the instructor's assessment of student performance.

The use of scaling

The scaling procedures are used in order to accomplish particular goals such as: a specific class average; boost poorer results in the test more than the better ones; boost better results in the test more than the poorer ones; how many students pass or fail the test; the number of As or Cs in the test, for instance. These may be specific goals, which should be achieved in a way that is not offensive to the class and that preserves evaluation standards.

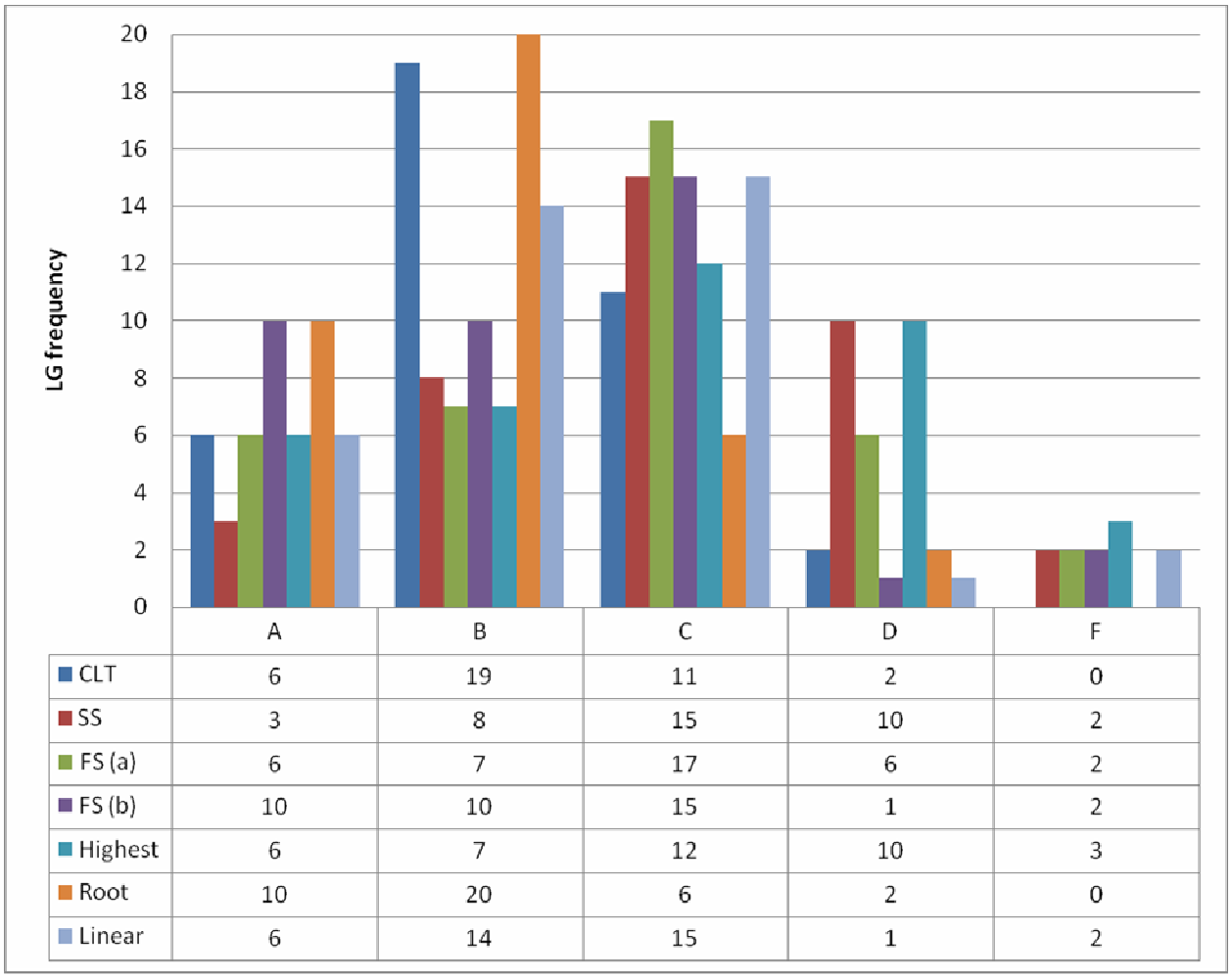
The scaling of raw scores can be performed up or down. The grade (as well as the raw score), however, has a psychological dimension to students. Scaling up is perceived by students as a bonus to their performance and scaling down as downgrading their performance. Accordingly, they are happier with scaling up procedures, while they are frustrated with scaling down procedures. Here is an example of such frustration:

"Does anyone else get ticked off by classes or departments where grades in a class must be scaled in order to ensure that certain people don't do as well as they otherwise would? Or, that TAs are essentially forced to give out bad grades, by, for example, being coerced by the administration to ensure that the class average is consistently at around 70% (a B- at my school), regardless of whether that is what the students in the class actually deserve? I find this grossly unfair as this has been happening with several of my classes…There was a person who got an 85% (A) brought down to a 73% (B) because of grading policies employed".[14]

The raw scores have their own impact on students. A very low raw score, even if scaled to reasonable level (from the student perspective), makes him/her happy but it does show the student that his/her performance was poor relative to what he/she could have done and it will likely be perceived as painful. Here is a student's reaction to excessive scaling:

"I dislike classes more when they have to scale so much; it's ridiculous. I'm in a class now which had a 35 average for a test. Scaling a 35 to an 85 (B centered) is just weird. Happiest I've been with a 46 ever." [14]

The impact of such excessive scaling may lower SET scores, which is certainly not favorable to the new faculty at many of the higher education institutions. These examples show that scaling may be useful for the fine tuning of grades, but it must be performed within students' tolerable limits, meaning in fact well calibrated tests for the material taught. Either very easy or very difficult tests will likely be detrimental to both student learning and to faculty. Therefore, scaling methods cannot by any means replace the instructor's close understanding of the students' problem solving potential.



	A	B	C	D	F
CLT	6	19	11	2	0
SS	3	8	15	10	2
FS (a)	6	7	17	6	2
FS (b)	10	10	15	1	2
Highest	6	7	12	10	3
Root	10	20	6	2	0
Linear	6	14	15	1	2

Figure 3. Comparison of LG distribution obtained using the 7 scaling methods discussed. The figure shows the frequency of LG categories corresponding to the scaling of raw scores from Table 1.

Figure 3 shows the LG distribution obtained by using the scaling methods discussed here, starting from the raw scores given in Table 1. Figure 4 also shows the class averages corresponding to scaled set. It is obvious that significant variation of LG distribution is present. Depending on the instructor's goal, various scaling scenarios can be chosen.

Table 1. Sample raw score data set.

95.652	91.304	91.304	86.956	86.956	86.956
82.608	82.608	82.608	82.608	78.261	78.260
78.260	73.913	73.913	73.913	73.913	73.913
73.913	73.913	69.565	69.565	69.565	69.565
69.565	65.217	65.217	65.217	65.217	65.217
60.869	60.869	60.869	60.869	60.869	56.521
47.8260	47.826				

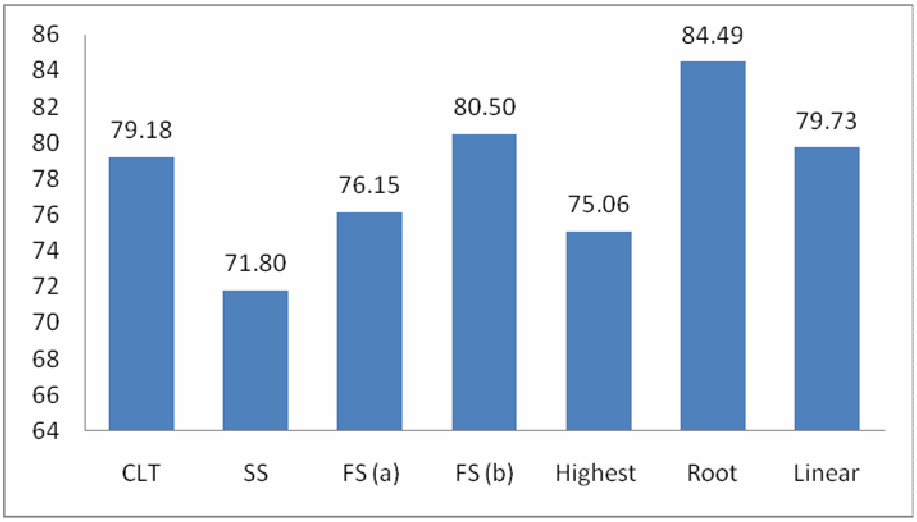


Fig. 4 Class average after using the scaling methods discussed (raw scores are given in Table 1). The vertical axis shows class average after scaling (%).

Conclusions

The scaling method of choice will definitely bear upon the results of grading and the meaning of the assessment of student learning, with real consequences for the students. Not knowing the effect of scaling in the grade distribution may be detrimental to both students and instructors. In order to alleviate the potentially harmful impact of grading on the student, the instructor's educated choices for scaling are needed. The real concern regarding scaling is the preservation of the meaning of the scaled score according to the standards recognized by the instructor. Scaling up is always preferable to scaling down, which brings about negative student perception. Therefore, it seems preferable to have the test more challenging always (however, not too challenging), as opposed to giving too easy tests. The intention of scaling is not to increase artificially the average of the group but rather to reflect group performance relative to objectives, difficulty, and other test variables. Diverse methods of scaling can be used; in our opinion, the most versatile one is the CLT, from among the ones discussed. The freedom of choosing LGs is still limited by the objective appreciation of test performance and by the acknowledgement that tests should be calibrated to class level so as not to appear offensive. Keeping this balance may ensure the coexistence of challenging tests with good SET scores (SET scores are obviously not exclusively the result of good grading).

References

[1] Ieta, A., R. Manseur, and T.E. Doyle. (June 14 – 17, 2009) "Effective criteria for teaching and learning." *The 2009 ASEE Annual Conference & Exposition,* Austin, TX, 2009.

[2] S. S. Stevens, "On the averaging of data," *Science*, Vol. 121, Jan. 1955, pp. 113–116.

[3] Wilbert J. McKeachie, and Marilla Svinicki. *McKeachie's Tips: Strategies, Research, and Theory for College and University Teachers,* Boston, Houghton, 2006.

[4] Lang, J. W. B., and M. Kersting. "Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run?" *Instructional Science*, vol. 35, nr. 3, May 2007. 187-205.

[5] McPherson, M. A., and R. T. Jewell. "Leveling the playing field: Should student evaluation scores be adjusted?" *Social Science Quarterly*, vol. 88, nr. 3, Sept. 2007. 868-881.

[6] Aigner, D. J. "On Student Evaluation of Teaching Ability." *Journal of Economic Education*, vol. 17, 1986. 243.

[7] McPherson, M. A. "Determinants of how students evaluate teachers." *Journal of Economic Education*, vol. 37, nr. 1, 2006. 3-20.

[8] Krautmann, A. C., and W. Sander. "Grades and student evaluations of teachers." *Economics of Education Review*, vol. 18, nr. 1, Feb. 1999. 59-63.

[9] Clayson, D. E., T. F. Frost, and M. J. Sheffet. "Grades and the student evaluation of instruction: A test of the reciprocity effect." *Academy of Management Learning & Education*, vol. 5, nr. 1, Mar. 2006. 52-65.

[10] Smallwood, M. L. *An Historical Study of Examinations and Grading Systems in Early American Universities*, vol. 24 of *Harvard Studies in Education*, New York, Johnson Reprint Corporation, 1969 (rpt 1935).

[11] Lissitz, R. W. and M. L. Bourque, "Reporting NAEP results using standards," *Educational Measurement: Issues and Practice*, Vol. 14, No. 2, 1995, pp. 14–23.

[12] Karran, T. "Pan-European Grading Scales: Lessons from National Systems and the ECTS," *Higher Education in Europe*, Vol. 30, No. 1, April 2005, pp. 5-22.

[13] Ieta, A., G. Silberberg, Z. Kucerovsky, and W. D. Greason, "On scales and decision-making based on arithmetic mean," *Quality & Quantity,* Vol. 38, No. 5, 2005, pp. 559–575.

[14] Scaling grades et al, http://talk.collegeconfidential.com/college-life/325006-scaling-grades-et-al.html

[15] Richeson, D. "How to curve an exam and assign grades", http://divisbyzero.com/2008/12/22/how-to-curve-an-exam-and-assign-grades/

[16] Ieta, A., Doyle, T. E., Kucerovsky, Z., and Greason, W. D. "Challenges and Options Related to Scaling Raw Scores in Engineering Education," The International Network for Engineering Education and Research. *Innovations 2009: World Innovations in Engineering Education and Research,* iNEER, Arlington (July 2009) Chapter 18, 219-234