Improving the Cross-Comparison of Educational Achievement and Evaluation

Adrian Ieta[1], Rodica Ieta[2], Thomas E. Doyle[3]

Abstract - **The achievements of engineering students are evaluated using particular grading scales. Arithmetic mean is usually employed for grade aggregation. We show that this measure renders correct results only for equivalent scales as defined in the paper. If grades from non-equivalent scales are aggregated, rank "errors" as well as "absurd" averaging may happen, as we originally observed in practice. Decision-making based on arithmetic mean aggregation of grades may be true, false, or fuzzy, according to our analysis. We also argue for the choice of grading scales in conducting cross-comparison of students' achievements. Our analysis also has relevance for regular grading and scaling methods, which we tested on engineering students with excellent student feedback. The grading technique proposed in this paper is particularly suited to engineering courses and it appears fit for promoting higher teaching and evaluation standards, paralleled by increased interest and higher competition levels for all student categories.**

Index Terms - achievement and evaluation, decision-making, equivalent scales, grading scales, grade or mark conversion.

INTRODUCTION

Internationalization of universities, mobility of students between departments, schools, states, or countries, distance education and globalization trends in general are factors contributing to engineering graduates having grades assigned on multiple grading scales. The problem of assessing student overall performance is obvious and arises from the need to compare achievements. Finding the right measure for this purpose is not the topic of our investigation here. Instead, we focus our attention on a measure often used in practice, namely, the arithmetic mean of students' grades and on how this can consistently be used for such a purpose.

Evaluating engineering students' course performance (but not only) may be a difficult task as marks have an important function for student selection processes [1]. Grading scales with different numerical assignments are used nationally and internationally for achieving this goal. Scales can be classified into different categories [2]-[5]. Different marking techniques try to measure student achievement and to group students in ordered categories, which actually do not have a linear correspondence to the degree of knowledge measured. Therefore, it is worth noting that the types of grading scales used in schools are ordinal [6, p. 114], [**7,** p. 305].

Letter grade (l.g.) scales represent an important type of grading scales; initially developed at Harvard [8], they now carry an intuitive meaning [9, p. 26] and are used extensively in North America. Although grades may carry some degree of subjectivity, they appear in transcripts as error-free.

Although grading itself is a difficult task, assessing overall performance in a variety of courses for reliable cross-comparison of student achievement is an even more challenging attempt. Whether more or less appropriate, the (weighted) arithmetic mean of individual grades is often employed in practice as a cross-comparison measure (it is well known that arithmetic mean is not an appropriate statistic for an ordinal scale [3], [6], [10], and educational grading scales are in fact ordinal). Some decision-making processes are based on the interpretation of the (weighted) arithmetic mean of measured data sets and are often employed in schools for candidate selection. In a recent work [10] we proved a theorem for the selection of ordinal scales that are equivalent from the point of view of arithmetic mean. In this case, comparison of arithmetic mean (ranks) originating on different ordinal scales is appropriate.

Some essential properties of particular grading scales used in conjunction with the arithmetic mean will be presented in this paper by means of examples. We are not dealing with mathematical proofs for the statements we make. The mathematically inclined reader is directed to [10], [11] for such purpose. Although letter grades will be used in our examples, this does not significantly limit the generality of our analysis of scales, as l.g.s will stand for common ranks with a particular meaning associated with student achievement. Using (weighted) grade averages [12] in assessing overall student achievement has certain consequences, not necessarily intuitive. The result of mathematical manipulations of grades as frozen measurements can be evaluated and it constitutes the reason for presenting our research results.

Our interest in this topic was prompted by several counterintuitive examples we have come across. For instance, one graduate student with a degree from a reputable university in North America started a graduate program at another reputable university in North America. His course averages were A- and A, respectively, on the grading scales of the

[1] Adrian Ieta, Murray State University, Murray, KY, adrian.ieta@murraystate.edu
[2] Rodica Ieta, Murray State University, Murray, KY, rodica.ieta@murraystate.edu
[3] Thomas E. Doyle, McMaster University, Hamilton, ON, Canada, tdoyle@ieee.org

issuing institutions. The second institution required an overall A- average in order to be admitted in scholarship competitions; although the student's achievements have been ranked as at least A- for each individual school, his overall calculated average (for both schools) has never met the minimum A- criterion to qualify for scholarships.

The goal of this paper is to contribute to the deeper understanding of results observed in practice (such as the one mentioned above) and to provide support for improving decision-making procedures relying on grade averaging as an overall assessment of student achievement. Based on the criteria we developed for equivalent grading scales, we suggest marking procedures when scaling/curving of the grades is involved.

LETTER GRADES AND ARITHMETIC MEAN AS AN OVERALL MEASURE OF STUDENT ACHIEVEMENT

The letter grade scale can usually be represented as a set

$$S = \{A, B, C, D, E, F\}$$

where the letter grade categories are associated to hierarchic ranks A >B > C> D > E > F (" >" stands for "higher in rank than") and individual letter grades may have subranks (for instance A+ >A > A− > B+ > B > B− > C+ >C > C− > D+ > D > D−). For the purpose of our examples, we will consider the following l.g. (sub)scale: S={A+, A, A−, B+, B, B−, C+, C, C−}. A popular numerical assignment associated with this l.g. scale is S_a={A+=4.(3), A=4, A−=3.(6), B+=3.(3), B=3, B−=2.(6), C+=2.(3), C=2, C−=1.(6)}. Another numerical scale is often used in correlation with conversion of l.g.: S_b={A+=12, A=11, A−=10, B+=9, B=8, B−=7, C+=6, C=5, C−=4}. Assuming that measuring student performance resulted into a set of grades s^1 = {A, B, C} and the arithmetic mean of the numerical assignments needs to be calculated, this average is regularly employed as a measure of the overall achievement of the student. For instance, if the student was graded on scale S_a, then the set of grades is s_a^1 = {x_1 = 4, x_2 = 3, x_3 = 2} with an arithmetic mean of

$$M_a(s_a^1) = \frac{1}{r} = \sum_{j=1}^{r} x_j = \frac{4+3+2}{3} = 3. \qquad (1)$$

The common interpretation of this average is an overall educational achievement of rank B. Let the set of grades of another student be s^2 = {A, A, B, C} (corresponding to s_a^2 = {4, 4, 3, 2}). The average on scale S_a for this set of grades is

$$M_b(s_a^2) = \frac{4+4+3+2}{4} = 3.25 \ (< 3.(3) = B+).$$

The interpretation of this average is also a rank B (with the common assumption that in the case of an average falling in between the scales' numerical assignments, the rank of the average is associated with the lower l.g. of the numerical values). It can be easily verified that had the students been graded and evaluated on the S_b scale, results similar to those on S_a would be obtained for the rank of arithmetic mean: s_b^1 = {11, 8, 5}, s_b^2 = {11, 11, 8, 5} and

$$M_b(s_b^1) = \frac{11+8+5}{4} = 8 \ \rightarrow \text{rank B on } S_b;$$

$$M_b(s_b^2) = \frac{11+11+8+5}{4} = 8.75 \ (< 9) \rightarrow \text{rank B on } S_b.$$

Let us consider the scale S_c={A+=30, A=26, A−=22, B+=18, B=15, B−=12, C+=9, C=7, C−=5}. The l.g. set s^2 has the numerical form s_c^2 = {26, 26, 15, 7} with an arithmetic mean

$$M_c(s_c^2) = \frac{26+26+15+7}{4} = 18.5 \quad (< 22) \rightarrow \text{rank B+on } S_c.$$

Thus, evaluating the same set of l.g.s on scales S_a and S_b provided the same rank for the arithmetic mean (B) and a different rank (B+) on S_c. This simple example shows that the corresponding rank of the arithmetic mean depends on the numerical assignments of the scale where the average is calculated. Naturally, some questions arise:

1) is the correspondence of the ranks of average associated with S_a and S_b always valid?

2) what are the general conditions that numerical assignments for scales need to meet so that evaluating the rank of the average on either scale may always preserve the rank?

Empirical verification of the ranks of calculated averages (for the same set of l.g.s) for numerical assignments on both S_a and S_b does confirm that the two numerical scales are equivalent in this respect. We answered the second question by demonstrating a theorem for equivalent scales [10]. Our theorem says that if two numerical assignments of the same categories of a grading scale (for instance {A+, A, A-, B+, B, B-, C+, C, C-}) are linearly related, then the rank of the arithmetic mean of an arbitrary set of grades is the same, irrespective of the numerical scale. Let $x_1, x_2, \ldots, x_n$ be the numerical assignments increasing in values (and ranks) of a grading scale S_x = {$x_1, x_2, \ldots, x_n$}, and $y_1, y_2, \ldots, y_n$ the corresponding values (and ranks) for the grading scales S_y = {$y_1, y_2, \ldots, y_n$}. The two scales will produce the same type of ranks of the arithmetic mean for any set of gades if and only if

$$y_i = a\, x_i + b \quad (i = 1, 2, \ldots, n) \qquad (2)$$

where a and b are real constants. It can be verified that S_a and S_b are related according to relation (2), while S_a and S_c or S_b and S_c are not. Therefore, S_a and S_b are equivalent and they are not equivalent to S_c. If a certain group of students has been graded using numerical assignments from S_a, it would be correct to use S_b for comparison of grades or averages.

OVERALL STUDENT ACHIEVEMENT ASSESSED FOR SETS OF GRADES ORIGINATING ON NON-EQUIVALENT SCALES

It was previously argued that ranks of the same set of l.g.s may vary and we will call these scales non-equivalent. Any two scales with numerical assignments (ranks) not following a linear relation (2) are non-equivalent. However, even if such a linear relation existed, it would not render the scales equivalent unless a one-to-one correspondence of ordered ranks (and subranks) were present. For example, S_{a1}={A=4, A−=3.(6), B+=3.(3), B=3, B−=2.(6), C+=2.(3), C=2, C−=1.(6)} is not equivalent to S_a because there is no one-to-one correspondence for the ranks of the scales.

TABLE I
DIFFERENT NUMERICAL ASSIGNMENTS OF LETTER GRADE CLASSES

L_i (l.g.)	S_h	S_a	S_b
A+	90, …, 100	4.(3)	12
A	85, …, 89	4	11
A-	80, …, 84	3.(6)	10
B+	77, …, 79	3.(3)	9
B	74, …, 76	3	8
B-	70, …, 73	2.(6)	7
C+	67, …, 69	2.(3)	6
C	64, …, 66	2	5
C-	60, …, 63	1.(6)	4

Some numerical assignments for letter grade scales are given in table I for direct comparison.

It was argued that S_a and S_b are equivalent. There is a one-to-one correspondence between the ranks of S_h and those of S_a. However, S_h contains subranks having numerical assignments to be used, for instance, in calculating the average of grades. The subranks of S_h do not have distinct meaning in S_b and therefore S_h and S_a are non-equivalent with respect to arithmetic mean. A justified endeavor would be to assess the impact of calculating the overall rank of arithmetic mean of sets of grades when the grades originate on scales with nonequivalent numerical assignments. We will therefore study averages of combined sets of l.g.s from S_b and S_h. As previously argued, S_a and S_b are equivalent and therefore the results of the study are identical if we use S_a and S_h instead.

Let us consider a student who obtained a set of l.g.s s_b^1 on S_b (which becomes s_h^1 on S_h) and another set s_h^2 on S_h (which becomes s_b^2 on S_b). If the student's numerical grades obtained on S_h are $s_h^1 = \{100, 89, 89, 79, 66\} = \{ 100, 2x89, 79, 66 \}_h$ and those on S_b are $s_b^1 = \{11, 11, 10, 10, 9 \} = \{ 2x11, 2x10, 10, 9 \}_b$ then the calculation of the arithmetic mean for each set of grades gives

$M_h(s_h^1) = 84.6$ (< 85) → rank A- on S_h;

$M_b(s_b^2) = 10.2$ (< 11) → rank A- on S_b.

The pair of sets (s_h^1, s_b^2) characterizes overall student performance and if an overall arithmetic mean for all grades is to be calculated a conversion scheme/rule must be established. The usual conversion of sets of grades to/from S_h from/to numerical scale S_b is performed between similar l.g.s according to the intuitive correspondence indicated in Table I. In order to calculate the average of all grades on S_b, set s_h^1 needs to be converted to a set corresponding to s_b^1 on S_b. In order to perform this operation one has to identify l.g. classes $s_h^1 = \{A+, 2xA, B+, C\}$ with corresponding numerical values on S_b $s_h^1 = \{ 12, 2x11, 8, 5 \}_b$.

TABLE II
EXAMPLES OF ABSURD AVERAGING

$\{ M_h(s_h^1)$,	$M_b(s_b^2) \}$	→	$M_b(s_b^{1,2}) = M_b(s_b^t)$
{ A+ ,	A- }	→	B+
{ A ,	A }	→	B+
{ A- ,	A }	→	B+
{ A ,	A- }	→	B+
{ A- ,	A- }	→	B+
{ A- ,	A- }	→	B
{ B- ,	B- }	→	C+
{ B ,	B- }	→	C+

The $s_b^{1,2}$ concatenated set of grades s_b^1 and s_b^2 gives

$s_b^t = s_b^{1,2} = \{ 12, 4x11, 2x10, 9, 8, 5 \}_b$

and its average is

$M_b(s_b^t) = 9.9$ (< 10) → rank B+ on S_b;

The following counterintuitive result has been obtained: the average of two sets of grades, both with average ranks of *A*−, resulted into a concatenated set of converted grades with an arithmetic mean of rank *B*+

$$\{(A-)_h, (A-)_b)\} \rightarrow (B+)_b .$$

Many other examples of this type can be found, as shown in Table II.

One would expect that the rank of the arithmetic mean of two concatenated sets would always lie between the ranks of averages of each independent set [10]. The explanation for this apparent contradiction is rooted in the fact that for all examples corresponding to cases in Table II, the rank of average for s_h^1 (when converted to S_b) is always lower than the one calculated for the concatenated sets $s_b^{1,2}$:

$$M_b(s_b^1) < M_b(s_b^{1,2}) = M_b(s_b^t) .$$

For instance, in the case of the numerical example discussed above, the set of grades $s_h^1 = \{ 100, 2x89, 79, 66 \}_h$ with an A-rank of average on S_h becomes $s_b^1 = \{ 12, 2x11, 8, 5 \}_b$, when converted to S_b, with a B+ average. Therefore, from a set of grades of B+ rank and one of A- rank, it is possible (it is also common-sensical) to obtain a B+ rank for the average of the concatenated set of grades. If the arithmetic mean is to be calculated on S_h (as opposed to S_b), to the same l.g. from S_b would correspond multiple numerical values on S_h. Therefore, converting s_b^1 to S_h (and obtaining s_h^1) is fuzzy. However, minimum and maximum numerical values can be identified according to the minimum and maximum values associated with each l.g. class in S_h (see Table I). The fuzziness of grade conversion is further propagated into the calculations of arithmetic mean of concatenated sets $(s_h^1, s_h^2) = s_h^{1,2} = s_h^t$.

We calculated averages for all possible combinations of l.g. as given in Table III. The comparison of the ranks of arithmetic mean can be summarized as follows:

- 25 l.g. combinations (69.4%) - the rank on S_b is lower than or equal to the one on S_h;
- 1 l.g. combination (2.7%) - the rank on S_b is strictly lower than the one on S_h;
- 7 l.g. combinations (19.4 %) - the ranks on both scales are equal;
- 3 l.g. combinations (8.3 %) - the rank on S_b is higher than or equal to the one on S_h.

The rank of the concatenated sets of grades is not invariant to the scale but rather heavily depends on the inherent properties of the numerical scales. It can be concluded that calculating the average on S_b will normally result in ranks of the average lower than the corresponding ones calculated on S_h, scale S_b having a strong bias in this respect. Although the analysis here was performed on two particular non-equivalent scales, the result is rather general and the issue of scale selection is of utmost importance. If the rank of the average of sets of grades is considered to be correct on a particular scale, then other ranks that may be obtained on other scales can be considered erroneous. These rank "errors" may occur and they may consist of one, two, or more rank deviations when sets of grades from non-equivalent ordinal scales are compared.

CRITERION FOR CHOOSING A 'FAIR' SCALE FOR EVALUATING STUDENT ACHIEVEMENT

Assuming that a student selection process is based on a minimum arithmetic mean rank of the overall sets of grades, conversions and multiple non-equivalent scales are usually involved. Decision-making processes may be inadvertently influenced by the very choice of the scale used for overall student evaluation. Grading students' performance orders students in groups of hierarchic ranks for each individual course. Numerical assignments corresponding to each rank (for instance l.g.) are of no relevance as long as the hierarchy of ranks is known. However, evaluating overall student achievement may vary according to scale when non-equivalent scales are used.

The averaging properties of the scales are dictated by their numerical assignments. Thus, each school or department implicitly adheres to hierarchic values generated by the averaging properties of its chosen type of scale used for grading its students. With this observation in mind and given that in general students have sets of grades obtained on numerical scales that are non-equivalent to the scale of the school or department where the overall assessment takes place, the 'fair' scale to be used for such purpose is obviously the scale employed by the school or department where the evaluation is performed. Although the conversion of grades to that scale may not be unique (as in the example of conversion from S_b to S_h), it provides a proper indication of how students could have performed overall on that scale, which represents the hierarchic values applied in the case of the school's or department's own students. In such a case, the comparison would be consistent with the locally accepted hierarchic values, as it is performed uniformly for all students and all sets of grades.

TABLE III
AVERAGING ELEMENTARY LETTER GRADE SETS

l.g. elementary sets s^{ti} = { L_i , L_j }			$M_h(s^{ti})$ min	$M_h(s^{ti})$ max	$M_b(s^{ti})$	Rank $M_h(s^{ti})$ min	Rank $M_h(s^{ti})$ max	Rank $M_b(s^{ti})$
s^{t1}	A+	C-	75	82	8	B	A-	B
s^{t2}	A+	C	77	83	8.5	B+	A-	B
s^{t3}	A+	C+	79	85	9	B+	A-	B+
s^{t4}	A+	B-	80	87	9.5	A-	A	B+
s^{t5}	A+	B	82	88	10	A-	A	A-
s^{t6}	A+	B+	84	90	11	A-	A	A-
s^{t7}	A+	A-	85	92	11	A	A+	A
s^{t8}	A+	A	88	95	12	A	A+	A
s^{t9}	A	C-	73	76	7.5	B-	B	B-
s^{t10}	A	C	75	78	8	B	B+	B
s^{t11}	A	C+	76	79	8.5	B	B+	B
s^{t12}	A	B-	78	81	9	B+	A-	B+
s^{t13}	A	B	80	83	9.5	B+	A-	B+
s^{t14}	A	B+	81	84	10	A-	A-	A-
s^{t15}	A	A-	83	87	11	A-	A	A-
s^{t16}	A-	C-	70	74	7	B-	B-	B-
s^{t17}	A-	C	72	75	7.5	B-	B	B-
s^{t18}	A-	C+	74	77	8	B-	B	B
s^{t19}	A-	B-	75	79	8.5	B	B+	B
s^{t20}	A-	B	77	80	9	B+	A-	B+
s^{t21}	A-	B+	79	82	9.5	B+	A-	B+
s^{t22}	B+	C-	69	71	6.5	C+	B-	C+
s^{t23}	B+	C	71	73	7	B-	B-	B-
s^{t24}	B+	C+	72	74	7.5	B-	B	B-
s^{t25}	B+	B-	74	76	8	B-	B	B
s^{t26}	B+	B	76	78	8.5	B	B+	B
s^{t27}	B	C-	67	70	6	C+	C+	C+
s^{t28}	B	C	69	71	6.5	C+	B-	C+
s^{t29}	B	C+	71	73	7	B-	B-	B-
s^{t30}	B	B-	72	75	7.5	B-	B-	B-
s^{t31}	B-	C-	65	68	5.5	C	C+	C
s^{t32}	B-	C	67	70	6	C+	C+	C+
s^{t33}	B-	C+	69	71	6.5	C+	B-	C+
s^{t34}	C+	C-	64	66	5	C-	C	C
s^{t35}	C+	C	66	68	5.5	C	C+	C
s^{t36}	C	C-	62	65	4.5	C-	C	C-

FUZZINESS AND BIAS

Let us consider an engineering candidate selection process involving national and international students. If S_G is the scale locally employed for grading students, and S_G, S_i are two scales used for evaluation (S_G to be used for students having grades only from the school where the selection takes place, and S_i to be used when students also have grades assigned on other numerical scales), the candidates can be grouped in categories [11]:

$C1$ - candidates having grades only from the school where the selection takes place (scale S_G) and evaluated on S_G;
$C2$ - candidates having grades from the school where the selection takes place as well as from other scales equivalent to S_G and evaluated on S_i;
$C3$ - candidates having grades from the school where the selection takes place (non-equivalent to S_i) as well as from other scales equivalent to S_i and evaluated on S_i;
$C4$ - candidates having grades from the school where the selection takes place (scale S_G non-equivalent to S_i) as well as from other scales nonequivalent to S_i and evaluated on S_i.
In agreement with our analysis, a few remarks may be of interest:

- Students belonging to $C1$ will always be correctly assessed in terms of their rank of arithmetic mean as the scale chosen for assessment is the one used by the school where they belong.
- Converting all the grades of the students in group $C2$ to scale S_i leaves the rank of arithmetic mean unchanged as the grades of these students were obtained on scales equivalent to S_i; thus their rank assessment is consistent with the hierarchic ranks to which the school organizing the selection adheres.
- For students in $C3$ the conversion of grades from other scales (except S_G) to scale S_i is consistent with the rank assessment on those scales as they are equivalent to S_i. Although calculating the rank of the overall set of (converted) grades on S_i is appropriate with respect to grades obtained on scales other than S_G (all equivalent to S_i), it is not appropriate from the point of view of the hierarchic ranks to which the school making the selection adheres. As S_i is non-equivalent to S_G, different hierarchic ranks of the arithmetic mean are associated with it and therefore with students belonging to $C1$ and $C2$. Converting grades from S_G to S_i may also cause fuzziness of the overall assessed rank (see examples of grade conversion from S_b to S_h).
- For students belonging to $C4$ the conversion of all the grades from scales non-equivalent to S_i is not appropriate as it is not consistent with rank assessment on S_G, to which the school adheres. Also, converting grades to S_i may cause fuzziness of the overall assessed rank as discussed for category $C3$ students.
- If only category $C1$ and $C2$ are involved in the selection process, the rank identification is always correct, as argued above, and thus the decisions made are always true.
- If category $C3$ or $C4$ students are also involved in the selection process, rank identification may be unique or fuzzy; as the assessment of rank is obtained on scale S_i (non-equivalent to S_G) for category $C3$ or $C4$ students, the decision-making process is altered by the fact that different categories of students are assessed on scales with different hierarchical values: category $C3$ and $C4$ students are assessed on S_i while category $C1$ and $C2$ are assessed (essentially) on S_G. Accordingly, the decisions made in such cases may be true, false or fuzzy; if S_G and S_i are strongly biased in terms of averaging properties (see the example of S_b, which is strongly biased to provide lower ranks of arithmetic mean with respect to S_h for the same l.g. set.), this may result in consistently accepting or rejecting candidates from a particular school due to its grading scale [11] rather than to student achievement.

GRADING AND EQUIVALENT SCALES

Grading tests may consist in ranking student work on a particular numerical scale. Many times grades are not directly obtained on a scale but rather a raw score is processed for such purpose. Let us consider the following numerical scale S_e = {A = {90, . . ., 100}, B = {80, . . ., 89}, C = {70, . . ., 79}}. If for a small class the results of the tests produced a set of raw scores (out of 100 or percentage) s_{raw} = {98, 72, 70, 60, 51, 8} numerical grades need to be assigned on S_e. The fact that students did not necessarily obtain high scores does not mean that the entire class deserves low grades but rather reflects the level of difficulty of the test with respect to that class (compare a regular mathematics test to a test given at an international competition, for instance). Therefore, even scores of 50 or 48 may indicate a significant level of student achievement, depending on the nature of the test. Assigning numerical values to be used for further processing (averaging) is an important and non-trivial task. Given that the instructor is able to assess what raw score range should correspond to A, B, or C, the issue of converting the effective raw score set to the numerical assignments of A, B, C, etc., arises. If the instructor assesses that the l.g. and raw score ranges are related according to l.g. categories according to {A = {80, 81, . . ., 100}, B = {60, . . ., 79}, C = {45, . . ., 59}}, what would be a fair correspondence to the numerical values of S_e scale? From the point of view of the analysis of equivalent scales presented in this paper, we propose that a fair correspondence would be obtained by making the l.g. subscales equivalent (numerical values for the A range on S_e scale should be obtained by a linear transformation of the 'raw score' range for A, and so on for the rest of l.g.s.). With such a transformation, the numerical grades obtained from s_{raw} = {98, 72, 70, 60, 51, 48} is $s_e = \{99, 86, 85, 80, 74, 72\}_e$. The procedure is consistent in ordering students within the same l.g. category and in preserving the arithmetic mean subranks. We tested this grading/ scaling method on different engineering and non-engineering students and on variable class-sizes, always with an excellent feedback from students.

The curving technique was applied, for instance, to ETE264, ETE366, and ETE470, all electrical engineering courses. The class was informed that the graded tests would include two numbers: the raw score, obtained through a strict marking procedure, and another one, representing the grade on the chosen grading scale. The curved grades spanned the regular range of A, B, C, D, and E l.g.s. At the end of the course the students were asked to respond anonymously to a questionnaire and to give their sincere opinion regarding the fairness of the evaluation of their course performance. Although no student can be pleased with low grades, there was

almost unanimous agreement that the evaluation of performance in the course was fair; only in a few cases students considered that they probably obtained a better mark than expected. The raw score associated with their grades likely made them think that they may have deserved a lower grade. The reaction of these few students proves that some of the students will show inertia to accepting other hierarchic values than those they were used to. It is worth noting that the instructor who employs this technique may appear to students as both a tough and a generous marker (because of the scaling procedure). The curving method can also affect the type of questions used in tests. In a highly heterogeneous class, difficult questions will crash the weaker part of the class, while too easy questions will not differentiate among students in the upper levels of the class. The scaling method presented here allows for the use of test questions with an overall higher difficulty level than it would be possible without any scaling, without compromising the correct and consistent ranking of students. Such an opportunity obviously promotes higher engineering education standards. In our practice, we noticed that the technique particularly engaged more students in continuously improving their engineering skills and understanding of concepts, while it also enhanced class interest and competition at all levels.

Although the grade curving procedure described in this section may appear cumbersome to apply in practice, our experience shows that it becomes a trivial operation once a spreadsheet is set for this purpose.

CONCLUSIONS

Cross-comparison of student achievement as reflected in grades is often performed in practice employing weighted averages and grade conversion. When two or more non-equivalent scales are involved in the calculation of the arithmetic mean rank, the rank of the concatenated set of grades may be lower than those of each of its independent grade sets (on their original scale). We called this phenomenon "absurd" averaging and explained the result. We do not claim to have found the ultimate solution to assessing overall student achievement irrespective of the grading scale used. Nor do we argue for the appropriateness of using the arithmetic mean for assessing overall student achievement. However, we do argue that when the arithmetic mean is used as such measure (with all its inherent problems related to the meaning of the result) the procedure is consistent only for equivalent scales (as defined in the paper). Also, when converting grades for the purpose of finding an overall parameter for student achievement reflected by sets of grades obtained on non-equivalent scales, for the sake of consistency, the conversion should be performed to the scale used by the student's current school or department. We showed how the equivalent scale analysis can be applied to regular grading procedures when raw scores are involved. Besides the theoretical aspect of this approach, our practice has consistently been paralleled by excellent student response to this new scaling/ curving procedure. The technique appears rewarding from the students' point of view and it is conceptually consistent with the analysis presented in this paper. In our opinion, the procedure is likely to be most efficient in allowing for enhanced teaching and evaluation standards, particularly in engineering courses, where exercises with higher levels of difficulty can be included in tests without affecting the overall efficiency of the evaluation. Although we have not solved the problem of overall student assessment, we provide insights based on mathematical grounds and suggest procedures meant to improve this process.

REFERENCES

[1] Natriello, G., "Marking Systems," *Encyclopedia of Educational Research*, vol. 3, sixth edition, Chicago: Macmillan, 1992, 772–776.

[2] Stevens, S. S., "On the theory of scales of measurement," *Science*, vol. 103, June 1946, 667–80.

[3] Stevens, S. S., "Mathematics, Measurement, and Psychophysics," *Handbook of Experimental Psychology*, New York: Wiley, 1951, 1–49.

[4] Coombs, C. H., "Psychological scaling without a unit of measurement," *Psychological Review*, vol. 57, 1950, 145–58.

[5] Torgerson, W. S., *Theories and Methods of Scaling*, London: Wiley, 1958.

[6] Stevens, S. S., "On the averaging of data," *Science*, vol. 121, Jan. 1955, 113–6.

[7] Khurshid, A., and H. Sahai, "Scales of Measurements: An Introduction and a Selected Bibliography," *Quality & Quantity*, vol. 27, 1993, 303–24.

[8] 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).

[9] Lissitz, Robert W., and Marry Lyn Bourque, "Reporting NAEP results using standards," *Educational Measurement: Issues and Practice*, vol. 14, no. 2, Summer 1995, 14–23.

[10] 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, 559–75.

[11] Ieta, A., Z. Kucerovsky, W. D. Greason, and G. Silberberg, "Fuzziness and Bias in Decision-Making Processes Using an Arithmetic Mean Criterion," *Quality & Quantity*, vol. 40, no. 2, 2006, 145–56.

[12] Bouyssou, Denis, T. Marchant, M. Pirlot, P. Perny, A. Tsoukiàs, and P. Vincke, "Building and Aggregating Evaluations," *Evaluation and Decision Models: A Critical Perspective*, Boston: Kluwer Academic Publishers, 2000, 29–53