Research in Higher Education, Vol. 41, No. 2, 2000

JUSTICE AND TRUTH IN GRADES AND THEIR AVERAGES

John M. Vickers

Grade point averages (GPAs) are calculated by assigning numbers to letter grades and averaging them. Simple examples show that the method cannot consistently determine class rank since class rank is sometimes permuted with arbitrary change of scale. This permutation is only possible when one student is somewhere worse and somewhere better than a second. The distinction between these and other sorts of cases is established by theorems proved in an appendix. Relativistic attempts to resolve the inconsistency are shown to be insufficient. The function of GPAs as predictors is briefly discussed.

1. INTRODUCTION[1]

The function and use of grading in secondary schools, colleges, and universities, are much discussed, and understandably so, for academic grades have deep and lasting effects in people's lives. Although performance on specific examinations commonly serves as a criterion for judging the suitability of applicants for professional and graduate schools, undergraduate grades, typically through grade point averages (GPAs) also play a significant role in these decisions. The role of secondary school grades and their averages in admission to colleges and universities is even more pronounced; a difference of a few points in GPA may mean the difference between acceptance and rejection by the college of one's choice.

Grading is often criticized as overly subjective; the widespread phenomenon of grade inflation is lamented. A grade from a large public high school is worth less than the same grade from a top-echelon preparatory school. There is no agreement on what grades mean in general or in particular: The question of what is or should be measured is vexed. Is the object of grades student perfor-

John M. Vickers, Department of Philosophy, Claremont Graduate University, 736 North College Avenue, Claremont, CA 91711; e-mail: John.Vickers@cgu.edu.

0361-0365/00/0400-0141$18.00/0

mance, or aptitude, or effort, or teacher approval? Should grades take into account the extent to which a student's knowledge or performance improves in a course? Again, grading is applied in widely different ways. In some institutions both grades for a repeated course are counted, in others only the second grade. Extracurricular activities, such as volunteer work, though they have little or no bearing on students' knowledge or performance, can also affect grades. There is, in short, little uniformity in grading practices across and even within institutions.

Grade averaging, since grades are the raw material on which it works, is subject to these infirmities as well as others.[2] Grade averages, it is said, and one can hardly contest this, fail to distinguish easy from difficult courses; an A in pottery weighs as heavily in the GPA as does an A in calculus or in the philosophy of Kant. This has the undesirable effect of encouraging students to avoid difficult courses in favor of easy ones in the end of maximizing the GPA. Similarly, grade averages are insensitive to the number of courses taken; a student who takes three courses and does well in them will have a better GPA than her colleague who has the same grades in these same three courses and does less well in a fourth and more difficult course (Siegel and Anderson, 1991). Nor can grade averages distinguish between different skills: The historian who does poorly in physics has the same GPA as the physicist who does poorly in history, and neither of these is distinguished from the student who does mediocre work in both physics and history. But these differences in talent are essential to the evaluation of academic accomplishment and aptitude. There is also considerable confusion sown by the great diversity of systems for averaging grades. In efforts to overcome or neutralize some of the difficulties just mentioned, some institutions weight grades in difficult courses more heavily, either by using weighted averages or by giving bonus points for enrollment. Other institutions may add a factor to the GPA in the interests of affirmative action for minorities or underprivileged students.

The general function of grading and grade averaging is to preserve and transmit information. What is preserved and transmitted is necessarily a very incomplete abstraction from the total history of student performance. A student's grade in a course (one of the standard five A–E grades) is an effort to capture in one of five letters a relevant evaluation of her work in that course. A student's GPA abstracts even further; the GPA aims to carry information about a student's complete academic performance, in many courses, typically in three digits. Given the richness, the great diversity, and the complexity of academic work, it is not surprising that grades and grade averages are often insufficiently informative. It is highly implausible that any so sparsely structured system of information transmission could preserve and carry what we ask of grades and GPAs.

Although there is a voluminous literature on grades and averaging, the simple structural properties of the grading and averaging system have been largely ignored. And that is the topic of this article. In what follows, grading is simply

characterized as a system of measurement and grade averaging is taken to be a method—methods, for there are several ways in which the thing is done—for calculating averages or expectations of grades. The attention is focused on structural features. To simplify the exposition, and also to broaden the import of whatever results might be obtained, it is assumed throughout that grades provide an objective measure of the quality of student work, and that there is a clear meaning to the ordering of work in terms of quality; to comparisons of the form "work X is at least as good as work Y." Grades are assumed to be not relative to teachers or evaluators. These assumptions amount to assuming that the substantive difficulties with grading, such as those mentioned above, are not essential to the process and that the system could be improved so as to settle or obviate its shortcomings.

This simplification is a device to enable the examination of grade averaging unhindered by the aporias and infirmities of the grading process. The questions to ask about grade averaging are first, what its minimal essential functions are, and second whether present methods can fulfill those functions. The next section sets down some simple structural principles about grading and gives a few examples of numerical codings for representing letter grades. Section 3 shows that grade averaging is inconsistent. This is a partial and negative response to the question of whether it can fulfill what is required of it. This inconsistency is simple and obvious, so much so that it is surprising that it has not (as far as I have been able to find out) been pointed out before. Section 4 begins with a few general remarks on measurement and averaging. Two theorems (The Comparability and Incomparability theorems) then localize the inconsistency and distinguish cases in which averaging is inconsistent from those in which it is benign. Section 5 remarks briefly on the relativistic affection for inconsistent methods. Section 6 considers and solves a puzzle about GPAs, namely, how it is that an inconsistent method has nevertheless some (admittedly weak) predictive power. Section 7 discusses precision in grading and percentage grades. The final section is an effort to draw some general conclusions from the preceding results and to propose a few actions and policies. An appendix is included in which the Comparability and Incomparability theorems are proved.

Throughout, I have tried to make the exposition accessible and self-contained for a general audience.

2. PRELIMINARIES: FUNCTIONS AND PRINCIPLES OF GRADING

Before broaching the matter of grade averages, it will help to set down a few principles about grades themselves and to record some different ways of representing them numerically. Recall that we assume there is a clear objective meaning to the ordering of work in terms of quality; to comparisons of the form "work X is at least as good as work Y." A student's grade gives objective

information about the quality of her work in the subject. On these assumptions, work of any grade is superior to work of any lower grade. Grades thus determine relations of superiority ("X is better than Y") and equivalence ("X is equivalent to Y") that satisfy the following *simple ordering conditions*, for all work X, Y, and Z:

If X *is better than* Y and Y *is better than* Z then X *is better than* Z *(transitivity)*
If X *is better than* Y then Y *is* not *better than* X (*asymmetry*)
If X *is equivalent to* Y and Y *is equivalent to* Z then X *is equivalent to* Z (*transitivity*)
X *is equivalent to* X (*reflexivity*)
If X *is equivalent to* Y then Y *is equivalent to* X (*symmetry*)
Either X *is better than* Y or Y *is better than* X or X *is equivalent to* Y (*connexity*)

Conversely, in a sufficiently representative collection of work, partitioned or divided by these relations into five exclusive equivalence classes, satisfaction of these principles determines the grades of all work in the collection. (All the work that is better than all work to which it is not equivalent is A work, all the non-A work that is better than all non-A work to which it is not equivalent is B work, and so on.) To repeat, we assume here that grades and the corresponding ordering are objective; the relative value of the work is a genuine character of it. B work is in fact better than C work.

Under this assumption of objectivity, grades share the structure of the measurement of gem and mineral hardness as measured by the Moh's scale. This scale assigns numbers from 1 to 10 to minerals and gems. Hardness is determined by how difficult it is to scratch the substance: Talc, for example, has hardness 1, gypsum 2, calcite 3, flourite 4, apatite 5, orthoclasse 6, quartz 7, topaz 8, corundum 9, and diamond 10. As with grades and merit, the hardness of all gems and minerals is determined by an ordering in terms of relations *is harder than* and *is equal in hardness to* satisfying the same constraints as those listed above for value of academic work; whatever is harder than topaz is harder than quartz, and so on.

When it comes to assigning numbers to letter grades, there are a number of systems. Perhaps most common is a four-point scale depicted in Table 1.

TABLE 1. Four-Point Scale

A	B	C	D	F
4	3	2	1	0

TABLE 2. CGU Eight-Point Scale

A+	A	A–	B+	B	B–	C	U (=E)
8	7	6	5	4	3	1	0

Though this scale is widespread, other scales are also in frequent use. My institution until quite recently used the eight-point scale of Table 2, in which grades of D are not scored. This scale recently was replaced by a ramification of the four-point scale, depicted in Table 3, in which intermediate codings for plus and minus grades are interpolated. One even finds different scales used in the same institution at the same time. In one secondary school district, board policy requires ordering courses by difficulty in five levels.[3] Letter grades are assigned points as shown in Table 4 (Level I easiest, Level V most difficult).

Scale I of Table 4 is the four-point scale of Table 1. The eight-point scale of Table 2 differs from all five of the multiple scales in Table 4. To facilitate comparison, Table 5 gives the scale of Table 2 omitting plus and minus grades. Again, D is not scored. The above scales are displayed together in Table 6 for ease of reference:

Let's draw a few conclusions from these examples. First, the different scales all codify the same information, except for the omission of D grades in the CGU scales, and are easily translated among themselves without loss of information. From this it follows that differences among the scales are matters of convention and represent no facts about the merit of student work. That an A (= 4) has twice the value of a C (= 2) in the four-point and ramified four-point scales does not signify that A work is twice as good as C work, for this relation between the values of A and C holds in no other scale. That C (= 4) is midway between A (= 6) and D (= 2) in Scale III does not signify that C work is midway in merit between A work and D work, for this relation holds in no other scale. That is to say that comparisons of the form "X work is twice as good as Y work," and "X work is midway in merit between Y work and Z work" are not facts about the merit of student work; these are artifacts of the *conventions* that differentiate the scales. Indeed, the comparison of the above scales, all of which adequately represent letter grades, makes it evident that the simple ordering

TABLE 3. CGU Ramified Four-Point Scale

A+	A	A–	B+	B	B–	C	C–	U (=E)
4.0	4.0	3.7	3.3	3.0	2.7	2.0	1.7	0

TABLE 4. Multiple Scales

Level	A	B	C	D	E (=F)
I	4.00	3.00	2.00	1.00	0
II	5.00	4.00	3.00	1.50	0
III	6.00	5.00	4.00	2.00	0
IV	7.00	6.00	5.00	2.50	0
V	8.00	7.00	6.00	3.00	0

conditions describe completely the information carried in letter grades. This can be put succinctly: Any numerical assignment that assigns higher numbers to higher grades is an adequate representation. This modest conclusion is worth a definition and a principle:

An assignment γ of numbers to letter grades is an *ordinal scale* if and only if γ always assigns higher numbers to higher grades.

Ordering principle. All ordinal scales are adequate numerical representations of letter grades.

Comparisons among scales will be helped along by another simple definition:

A numerical function φ is a *monotone transformation* if and only if φ always assigns higher numbers to higher numbers.

Every monotone transformation of an ordinal scale is an ordinal scale, and all monotone transformations of any ordinal scale are adequate numerical representations.

3. AVERAGING: TRACKING THE SOURCES OF CONTRADICTIONS

Grade averaging is a simple procedure. Numbers are assigned to grades in some ordinal scale. The numbers assigned to certain grades, typically those of one student in different courses, are added and the sum divided by the number

TABLE 5. Partial CGU Eight-Point Scale

	A	B	C	D	E
8-pt	7	4	1	?	0

TABLE 6. Scales Compared

SCALE	A	B	C	D	E
Partial CGU Eight-Point	7.00	4.00	1.00	?	0
CGU Ramified Four-Point	4.00	3.00	2.00	?	0
I (Four-Point)	4.00	3.00	2.00	1.00	0
II (Five-Point)	5.00	4.00	3.00	1.50	0
III (Six-Point)	6.00	5.00	4.00	2.00	0
IV (Seven-Point)	7.00	6.00	5.00	2.50	0
V (Eight-Point)	8.00	7.00	6.00	3.00	0

of grades in question. The result is the *grade point average* (GPA) of the collection of grades or the student. The primary function of GPAs is obvious—they should provide a rank ordering of student work overall or on the average (Lang, 1997). In the simplest case, if two students have taken the same courses, the student with the higher GPA should have done better work overall. The work of students with the same GPA should be equivalent overall. Comparisons of students whose courses do not coincide will be more delicate, but the simple case will suffice for our purposes.

Grade averaging is sometimes benign and effective. If Aaron has five A's and Bernice has four A's and a C then Aaron has clearly done better work than Bernice overall. Their GPAs show this. Aaron has a higher GPA than Bernice no matter what ordinal scale is used as a basis of the calculation. (See Table 7.)

Other cases are however less clear. Suppose that Justine and Maurice each take 3 courses. Justine gets 2 A's and 1 E and Maurice gets 3 C's. Then Justine has the higher GPA in the CGU eight-point scale, the ramified four-point scale,

TABLE 7. Aaron and Bernice

Scale	Aaron's GPA (5 A's)	Bernice's GPA (4 A's, 1 C)
Parital CGU Eight-Point	7.0	5.8
CGU Ramified Four-Point	4.0	3.6
I (Four-Point)	4.0	3.6
II (Five-Point)	5.0	4.6
III (Six-Point)	6.0	5.6
IV (Seven-Point)	7.0	6.6
V (Eight-Point)	8.0	7.6

and Scales I and II (see Table 8). Maurice has the higher GPA in Scales IV and V, and they have the same GPA in Scale III. So 2 A's and one E are better, worse, or equivalent to 3 C's, depending on the scale. It follows from this that the relation between the average or overall merits of Maurice's and Justine's work is a matter of convention, not a truth about their work.

The source of these contradictions is a feature of ordinal scales remarked above, namely that comparisons of the form "X work is midway between Y work and Z work" are not facts about the merit of student work, but artifacts of the differences among adequate ordinal scales. We turn now to a more general consideration of this.

4. GRADE AVERAGING AS ORDINAL MEASUREMENT

The order of students' GPAs may vary with the ordinal scale that represents their grades. This means that GPAs cannot fulfill their primary function, which is to provide a rank ordering of student work overall or on average. The source of this is that midpoints are not invariant in ordinal measurement. A few general remarks about measurement should make it clear that this is essential to the method and not a matter of detail.[4]

Only some properties and relations of numbers represent properties and relations of objects or substances to which they are assigned in measurement. To assume otherwise—that every numerical property and relation represents some property or relation among the things numbered or measured—is a fallacy. Let us call this the *numerological fallacy*. The simple way to detect the numerological fallacy is to look at alternative systems of measurement that are equivalent in the sense that they carry the same information. Then relations that vary among these equivalent systems are artifacts of the process, not representative of reality. That is the method used above to show that grade midpoints are

TABLE 8. Justine and Maurice

Scale	Justine (2 A's, 1 E)	Maurice (3 C's)
Partial CGU Eight-Point	**4.67** >	1.00
CGU Ramified Four-Point	**2.67** >	2.00
I (Four-Point)	**2.67** >	2.00
II (Five-Point)	**3.33** >	3.00
III (Six-Point)	**4.00** =	**4.00**
IV (Seven-Point)	4.67 <	**5.00**
V (Eight-Point)	5.33 <	**6.00**

artifacts of the grading scale, not facts about the merit of student work. The measurement of hardness is analogous. That the integers from one to ten are used as labels for the hardness equivalence classes is immaterial; any numbers (e.g., 0, 1, 3, 4, 5, 6, 9, 11, 13, 15) that preserve the order of hardness will do as well. This lets us see that midpoints in assigned numbers do not correspond to any midpoint relation in hardness: Is it gypsum or is it orthoclasse that is midway in hardness between calcite and quartz? Five is midway between 3 and 7, 6 is midway between 3 and 9, so, since both numerical assignments adequately represent the facts, these differences correspond to no relation in fact. Midpoints in hardness do not exist. Of course the midpoints of the *numbers* that represent hardness exist, but this represents no objective feature of the objects or their relations. There are no substances that are midway in hardness between gypsum and orthoclasse, because "midway" has no sense in this context.

Measurement yields objective information strong to the extent that relations and properties do not vary with the scale. Thus weights and lengths support comparisons of the form "X is twice (or ten times) the weight or length of Y." We can see this by noticing that such claims are independent of, for example, avoirdupois or metric scales of weight: if X and Y weigh 1 and 2 kilos respectively, then they also weigh 2.2 and 4.4 pounds, and this ratio, X is half the weight of Y, must obtain in any adequate scale of weight measurement. The weight of one object, however, cannot be said to be the square of the weight of another, for this relation changes with the system: $1^2 = 1$ but $2.2^2 = 4.84$, so squares of weight do not exist.[5] Of course the squares of the *numbers* that represent weight exist, but this represents no objective feature of the weighed objects or their relations. That the avoirdupois weight of X is the square of that of Y refers to no property or relation of X and Y, it is a numerical property that represents no non-numerical fact.

Temperature provides other good examples of measurement strength. Temperature multiples do not exist: 5 Celsius equals 41 Fahrenheit; 10 Celsius equals 50 Fahrenheit. Ten is twice 5, but 50 is not twice 41. To say that today is twice as hot as yesterday is thus to subscribe to a numerological fallacy, importing into the physical phenomena irrelevant characteristics of the numbers used to represent them. Temperature measurement is—since temperature ratios are not objective—weaker than the measurement of weight or length.[6] It is, however, stronger than the measurement of hardness described in section 2 above, for temperature midpoints do exist. Celsius temperatures 5 , 10 , and 15 are equal to Fahrenheit temperatures 41 , 50 , and 59 . Ten is the midpoint of 5 and 15, just as 50 is the midpoint of 41 and 59.

The ordinal measurement of hardness and that of students' work, we have seen, is weaker still. Here neither the squares, nor multiples, nor midpoints of numbers assigned have any fixed meaning. It is only relations of order that are invariant with scale. These truths are summarized in Table 9.

TABLE 9. Invariants

Measurement	Greater than, Equal to	Midpoints	Multiples	Squares
Weight, length	yes	yes	yes	do not exist
Temperature	yes	yes	do not exist	do not exist
Hardness, grades	yes	do not exist	do not exist	do not exist

To say that midpoints are invariant is just to say that if γ and δ are adequate scales and X, Y, and Z any measured objects, then

$$1/2[\gamma(X) + \gamma(Y)] = \gamma(Z) \text{ if and only if } 1/2[\delta(X) + \delta(Y)] = \delta(Z)$$

Midpoints, we have just seen, are invariant in the measurement of weight, length, and temperature, but not in the measurement of hardness or in grading. C is midway between A and E in the four-point system, but just one-eighth of the way in the eight-point system. This is the precise source of the contradictions and relativity described in the previous section. Let us state this as a law about averages and midpoints:

Law of averages and midpoints. If averages exist in a form of measurement, then midpoints are invariant in that form.

Since midpoints do not exist in grading, averaging, as the examples of section 3 make evident, leads to inconsistencies. We saw earlier, however, that in some cases averaging is benign. This does not purge the method of its inconsistency, but it does invite a closer look at the two sorts of cases in an effort to localize that inconsistency.

A little vocabulary and a few principles will ease the way here. By a *profile* is meant a sequence of grades of fixed and finite length, k. Profiles are ordered k-tuples of grades. To keep things simple, we consider profiles all of the same length, k, and we assume as well that courses are of equal weight. We consider first the comparison of pairs of profiles, for the minimum essential function of GPAs is to provide an unequivocal comparison of these. Recall the examples of Aaron and Bernice and Maurice and Justine.

Aaron: <A, A, A, A, A>
Bernice: <C, A, A, A, A>
Maurice: <C, C, C>
Justine: <E, A, A>

Aaron is never worse and somewhere better than Bernice. It is easy to see and to prove that Aaron's GPA is higher than Bernice's in every ordinal scale: Suppose that a and c are the numbers that represent A and C respectively in an ordinal scale. Then $a > c$, so

$$5a > 4a + c$$
$$(1/5)(5a) > (1/5)(4a + c)$$

Conversely, Maurice has a higher GPA than Justine on the seven-point scale, in which $e = 0$, $c = 5$, and $a = 7$:

$$(1/3)(3c) = 5 > 4.67 = (1/3)(2a + e)$$

but Justine has a higher GPA than Maurice in the five-point scale (when $e = 0$, $c = 3$, and $a = 5$):

$$(1/3)(3c) = 3 < 3.33 = (1/3)(2a + e).$$

And they have the same GPA in, for example, the scale

$$e = 0,\ d = 2,\ c = 4,\ b = 5,\ a = 6$$

where

$$(1/3)(3c) = 4 = (1/3)(2a + e).$$

The comparative magnitude of Maurice's and Justine's GPAs is thus completely determined by convention, and it represents nothing about the comparative merit of their work.

Other comparisons are less obvious and require some manipulation to effect. Consider the pair p, q, of profiles

$$p = \langle C, D, A, D, E \rangle$$
$$q = \langle D, E, C, A, C \rangle.$$

Each is sometimes worse and sometimes better than the other. There are however permutations, p', q' of p and q in which p' is never worse and sometimes better than q':

$$p' = \langle E, D, C, C, A \rangle$$
$$q' = \langle E, D, D, C, A \rangle.$$

And reasoning as in the case of Aaron and Bernice shows that here too the average of p (= the average of p′) is higher in every ordinal scale than that of q (= the average of q′):

If e < d < c < a, then (1/5)(e + d + 2c + a) > (1/5)(e + 2d + c + a).

What counts, then, for the comparison of averages to be unequivocal is that there should be *permutations* of the profiles one of which is nowhere worse and somewhere better than the other. In such a case we say that the stronger profile *strictly dominates* the weaker:

p *strictly dominates* q if and only if there are permutations $p' =< p'_1, \ldots p'_k>$ and $q' =< q'_1, \ldots q'_k>$ of p and q such that $p'_i \geq q'_i$ for every i, and $p'_j > q'_j$ for some j.

Accompanying the definition of strict dominance is the obvious definition of *equivalence*.

profiles p and q are *equivalent* if and only if they have the same numbers of all grades.

Clearly p and q are equivalent if and only if for some permutations p′ and q′, $p'_i = q'_i$ for every i.

We say that p *simply dominates* q if p strictly dominates or is equivalent to q.

Strict dominance is a strict partial ordering of the set of all profiles; it is transitive, asymmetric, and irreflexive. Equivalence is an equivalence relation (transitive, reflexive, symmetric) and simple dominance is transitive and reflexive. Profiles are equivalent if and only if each simply dominates the other.

Let us say that profiles are *comparable* if either simply dominates the other. Similarly, a pair of profiles is comparable if the profiles are comparable. Though strict and simple dominance and equivalence are transitive, comparability is not: Maurice's and Justine's profiles <C, C, C> and <A, A, E> are both strictly dominated by and hence comparable with <A, A, A>, but they are incomparable with each other. Comparability classes must thus be defined in terms of *pairwise* comparability:

A class Γ of profiles is said to be a *comparability class* if every pair of profiles in Γ is comparable. If p and q are non-equivalent profiles in a comparability class then one strictly dominates the other.

A simple example of a comparability class is the class of all *constant* profiles—profiles all members of which are the same (the sequence of k A's, the sequence of k B's, etc.). Again, the class consisting of all profiles (of length k) which include some or no A's and some or no B's is a comparability class.

There are many overlapping comparability classes, and comparable profiles will not in general belong to all the same comparability classes. Each comparability class is strictly ordered by strict dominance: strict dominance is transitive and asymmetric, and if p and q are nonequivalent members of the same comparability class then one strictly dominates the other. So within a comparability class the ordering of profiles by dominance is unequivocal. We need assume only that strict dominance and equivalence represent difference and equivalence in merit to conclude that each comparability class is unequivocally ordered by relative merit:

Ordering assumption. If p strictly dominates q then p represents better work overall than q. Equivalent profiles represent work that is equal overall in merit.

Further, it now follows that grade averages within a comparability class are unequivocal and conform to dominance in a precise sense:

Comparability theorem. If Γ is any comparability class, γ is any ordinal scale, and p and q any members of Γ, then, where
$(p) = (1/k)\Sigma_i\ \gamma(p_i)$ and $(q) = (1/k)\Sigma_i\ \gamma(q_i)$
$(p) >\ (q)$ if and only if p strictly dominates q
$(p) =\ (q)$ if and only if p and q are equivalent
$(p) \geq\ (q)$ if and only if p simply dominates q

The simple proof is given in the appendix.

It follows from the comparability theorem and the ordering assumption that GPAs give an unequivocal representation of academic merit (as defined by the ordering assumption) within each comparability class.

The comparability theorem gives a sufficient condition for the applicability of grade averages. This condition is in fact also necessary. It is proved in the appendix (The Incomparability Theorem) that *if p and q are not comparable then there are ordinal scales in which p has a greater average than q, in which q has a greater average than p, and in which their averages are equal*.[7] It follows from this that the relative magnitude of GPAs between incomparable profiles does not represent dominance, or anything else for that matter. It depends completely on scale. This localizes the inconsistency in grade averaging.

The general situation is as follows. An inclusive and unrestricted collection of profiles (of fixed length k) will include numerous distinct comparability classes. Within each of these classes, rank in order of overall scholarly merit is determined by dominance and hence by GPA. In classes of profiles in which not all members are comparable, however, this ordering is perturbed. Although order among members of a comparability class is fixed by dominance and GPA,

order between these and profiles with which they are incomparable is arbitrary; dominance is not defined here and the order of averages varies with scale. This means that a profile may rise or drop in relative rank in class with arbitrary change of scale. A profile may be third in one comparability class, tenth in another comparability class, and it may shift between fourth and fifteenth in the inclusive class of all profiles, depending upon scale.

About this two remarks:

1. The numerological fallacy beckons here: The comparability theorem says just that the relations *greater than* and *equal to* among GPAs represent *strictly dominates* and *is equivalent to* among profiles within the same comparability class. The theorem justifies no comparisons of proportion or interval, even among members of the same comparability class, for if p strictly dominates q we can make the difference and ratio of their GPAs as (finitely and positively) great or small as we like by judicious choice of scale. Thus the relative sizes of differences in GPAs even among comparable profiles signify nothing.
2. Information decreases with each successive representation. In the beginning, let us suppose there is the complete record of each student's work and performance in each course throughout a scholarly career in, say, secondary school. This presents an enormous dossier, much too big to be comprehended as such, much less to be compared with those of other students. This huge corpus is first represented in the course titles and grades for each student. Absent the titles (and assuming for simplicity equal importance of courses), we are left with the profiles. So far the information, if sparse, is coherent. Comparison of comparable profiles allows a summary comparison of academic merit. Incomparable profiles permit no such comparison.

   Enter GPAs accompanied by the definition of the scale in which the grades are coded. If the collection of profiles forms a comparability class, and if we know this, then the information provided by GPAs is, though sparse and subject to the infirmities mentioned in the introduction, nevertheless coherent. Higher GPA within the comparability class represents strict dominance among profiles in that class, and by the ordering assumption, this represents better scholarship overall. But if, as is virtually always the case, the collection of profiles is heterogeneous, including distinct comparability classes and many incomparable profiles, then each GPA will represent a profile that belongs to several distinct comparability classes, so each will be in several distinct dominance orders. Some pairs of profiles will be incomparable, and the order of the corresponding GPAs will thus not represent dominance or a difference in scholarly merit. Other pairs will be comparable, and here the order of the GPAs will represent dominance and a difference in merit, but we can have no way of knowing which comparisons belong to which sort.

That information is lost in the abstraction from profile to GPA. There is no way to recoup it from the unadorned GPA, so the list of these carries more noise than information.

5. RELATIVISM AND CONVENTION

This merits mention only because several people whom I should not otherwise have thought differently abled apparently found it persuasive. If all parties used one adequate system, then anomalies like the reversal of Justine's and Maurice's standings might not come to light and the numerological fallacy might fester undetected. Grade point averages, the argument runs, indicate scholarly merit so long as some ordinal scale is universal. Ignorance, that is to say, is blissful freedom from what it does not know. The same bucolic reasoning would have it that, since the Moh's scale is the only measure of relative hardness in current use, gemologists could in fact say with impunity that orthoclasse is midway between gypsum and diamond in hardness. And if all reference to Celsius and Réaumur scales were suppressed, then it would in fact be twice as hot in Dallas (50) as in Great Falls (25). The fallacy here is patent: On the four-point scale C (= 2.0) is midway between A (= 4.0) and E (= 0). If that is correct, then Justine is a better student overall than Maurice, and the eight-point scale—according to which Maurice is a better student overall than Justine—is wrong, for there the interval between A (= 8) and C (= 6) is 1/4 of that between A and E (= 0). But both scales clearly and adequately represent all the facts. The choice between them is a matter of convention, like the choice of Fahrenheit or Celsius temperature and that of metric or avoirdupois weight measure. Whatever is a fact, as distinct from an artifact of the system of representing the facts, though it may be hidden, cannot be changed by even the most Draconian enforcement of convention. Since the relative overall scholarly merit of Justine's and Maurice's work varies with convention, there is no fact of the matter. That Justine is a better student overall than Maurice is not a fact present in or entailed by their scholastic records: 2 A's and an E versus 3 C's. There are no grade midpoints, and where there are no midpoints, there can be no averages.

One might try to save averaging by resorting to subjectivism about grades; to holding that grades are no more than the expression of the evaluator's opinion, that objective reference to quality is filtered through the appreciation of the evaluator. This is an appealing view, conforming as it does to what is known about the relativity of grades to evaluators, but it does nothing to strengthen measurement as would be required to support consistent averaging. It is a fact about the structure of grading that grades provide just a simple ordering, no more. Such an ordering yields numbers under the constraint that greater numbers represent higher places in the ordering, and this is insufficient to support averages. Evaluators' opinions are in this respect like individual preferences for

commodities or events. One gets preference midpoints only by some additional device such as (in the case of preference orderings) gambles: a 50-50 gamble on X and Y marks the preference or value midpoint between X and Y. There looks to be no analogous device for grading.

6. A PUZZLE ABOUT PREDICTION[8]

If grade averages are meaningless, how is it that they predict, at least to some extent, academic performance? Why do high school graduates with higher GPAs tend to have higher college and university grades? The answer to this question is pretty obvious, but worth giving since at least some people profess to be puzzled by it.

Assuming the objectivity of grading, a difference in the magnitude of averages between comparable profiles represents a difference in the merit of the work, and one may presume, in the scholarly capacities of the students. It is thus plausible that GPA differences among comparable profiles of secondary school graduates are correlated with differences in college and university academic performance. Differences in averages of incomparable profiles, however, are no indication of academic capacity and will have no such correlation. Further, as mentioned in section 4, differences in the bare averages between two profiles give no hint as to whether the profiles in question are comparable, in which case the difference is significant, or incomparable, in which case it is not. Thus, in a heterogeneous population in which some pairs are comparable and others incomparable, there will be an ambiguity in the meanings of GPA differences. Some differences will indicate a difference in scholarly talent or proficiency, others will be independent of this. To put this in terms of probabilities, let p and q be the profiles of students P and Q, and let (p) and (q) be the averages of p and q in some ordinal scale. Then (Prob[X / Y] is the conditional probability of X given Y)

Prob[P has more talent than Q / p and q are comparable and (p) > (q)] is high

and

Prob[P has more talent than Q / p and q are incomparable and (p) > (q)] = Prob[P has more talent than Q]

From which it follows that, when P, Q are randomly chosen students from this heterogeneous population,

Prob[P has more talent than Q / (p) > (q)] > Prob[P has more talent than Q].

Hence, on the assumptions mentioned, given that (p) is higher than (q) and no other evidence, it is a better than even bet that P will do better than Q in college or university studies. It must however be kept in mind that this probability is a mix of probabilities of two quite distinct sorts; one sort is an indicator of academic talent, the other is independent of it.

Further, and worse, as the number k of courses increases the proportion of incomparable profiles increases very rapidly. Thus for a given student population, as the number of courses increases the second of the above conditional probabilities takes on greater weight and proportionally fewer differences in grade average will indicate difference in scholarly talent or proficiency. So, the more comprehensive and varied the data, the worse the correlation between grade average and what it is supposed to predict. That is the very definition of a non-robust method.

7. PRECISION AND PERCENTAGES

The difficulties with grade averaging do not have to do with precise versus vague ways of measuring. It is not a matter of figuring things out to a lot of decimal places, as the above discussion should make clear. Not every comparative concept admits of taking midpoints: Are potato chips midway in saltiness between chocolate ice cream and anchovies? Discriminations may be quite fine—as they are in some of us for saltiness—without determining midpoints. What would be required to make grade averaging meaningful are principles, not mere conventions, that would show this and why some ordinal scales are correct and others not. The principle that, for example, an A and an E should average to a B is obviously no help here: That is just to say that B should be the midpoint between A and E, and can hardly be advanced as support for itself.

Percentages, when used with care, are free of the vicissitudes wrought by changes of scale that infect grade points. The weighted average of percentages is under certain conditions again a percentage, for a percentage or proportion is an absolute measure. Whoever has taught knows that percentage calculations can often be a great help in grading when the conditions are right; when, for example, questions of equal and independent importance constitute an examination. But if conditions are not right, if the knowledge or skills to be mastered does not admit of quantification in this way, then to record a grade as a percentage rests upon a vitiatingly false presupposition: The quality of Louise's lucid and elegant proof of the theorem clearly exceeds that of Justine's more pedestrian if adequate argument. This does not mean that there is some material of which Louise has mastered 95% to Justine's 75%. These numbers signify no more than an ordering in this context, and their amalgamation in averages is hence equivocal and baseless, yet another resort to the numerological fallacy.

Further, percentages must be weighted in averaging: Differential grading poli-

cies are a fertile ground for confusion about this. Since "It [is] not reasonable to assume that grades from all courses at a university are determined by a single ability construct," researchers have used factor analysis to partition courses into subsets of comparable courses.[9] Suppose there are just two exclusive classes: natural science (NS) and humanities and social science (HS) and that Justine passed three of her four NS courses and failed the one course she took in HS. So she passed 75% of her NS courses and 0% of her HS courses. Maurice, on the other hand, passed his one NS course and just one of his four HS courses, so he passed 100% of his NS courses and 25% of his HS courses. (See Table 10.)

Maurice thus passed a higher percentage of courses than did Justine both in NS and in HS (100% to 75% and 25% to 0%). But Justine passed a higher percentage overall—60% to Maurice's 40%, and Maurice's (unweighted) average percentage of courses passed is the average of 100% and 25% or 62.5% which is higher than Justine's average of 37.5%. But Justine passed a higher percentage overall—60% to Maurice's 40%. The right calculation here is not the raw average of the percentages, but the average weighted by the proportions of courses in each term. Justine took 4/5 of her courses in NS and Maurice took 1/5 in NS. These ratios are reversed in HS. So the weighted averages of courses passed are:

Justine: $(4/5)(75\%) + (1/5)(0\%) = 60\%$
Maurice: $(1/5)(100\%) + (4/5)(25\%) = 40\%$

The moral is simple: Handle percentages with care.[10]

8. CONCLUSIONS AND MODEST PROPOSALS

A few conclusions are easy to state.

First, grades do not support consistent averaging. A difference in grade averages sometimes indicates a difference in scholarly accomplishment, when one profile strictly dominates another, but more often is an artifact of scale. Arbitrary or conventional shift of scale, independent of academic performance, can permute students' rank in class, and that is the source of the inconsistency. Since

TABLE 10. Weighting Percentages

Subset	Justine Passed		Maurice Passed	
NS	3/4	75%	1/1	100%
HS	0/1	0%	1/4	25%
All Courses	3/5	60%	2/5	40%

grade averages carry no pedigree, they do not distinguish between the two sorts of cases. This leads to an amalgamation of significant and insignificant numerical comparisons.

Grade averages do not predict performance very well, but it is a question why they should predict it at all, why a higher GPA lends any probability whatever to a prediction of scholarly superiority. The amalgamation of the two sorts of averages—significant and insignificant—answers this.

Is there one capacity, let us call it scholarly competence, present in varying degrees in the academic population, that manifests itself in scholarly performance? Probably not, but this article neither offers nor presumes a response to the question. If there is such a capacity, its extent will be most clearly manifest in cases of strict dominance, when one student's work is never worse and sometimes better than another's, and there grade averages tell the right story. When profiles are incomparable, the data are silent on the presence and comparative extent of the capacity. And we have seen that the use of grade averages to predict academic performance is precisely the contrary of a robust method: more data means worse fit.

It must also be emphasized that the contradiction in grade averaging depends not at all on the evident and often remarked difficulty of representing distinct capacities, such as those required in different subjects, in one index. Nowhere has it been presumed that different skills are at issue in the different elements of profiles. Indeed, it is compatible with the above argument that the elements of profiles are grades in the same subject. One can, for example, take profiles to give the grades of different students in the same course: So, suppose the five football players in a psychology course all get A's, and that the five basketball players get four A's and a C in the same course. Then the footballers are better psychologists than are the basketballers, and the GPAs of the two profiles show this. No question. If, however, the footballers get all C's and the basketballers three A's and two D's, then the data entail nothing about the relative psychological talents of footballers and basketballers. The contradiction in grade averaging is a structural feature of the method, dependent not at all upon substantive principles.

The argument to the conclusion that grade averaging is inconsistent assumes the simplest case: courses are symmetrical, none being more important than any other, all profiles include the same number of courses, and averages are direct and unweighted. Differential weighting or otherwise tinkering with the method cannot soften the inconsistency, for the simplest case will always be a special instance of its ramifications.

Nothing that is said here should be construed as an argument against grading. Grading is an obvious and essential part of pedagogy. Students have a right to fair and competent evaluation, and most who know about the matter will agree that teachers and evaluators do all that they can to assure that this right is

respected. Certainly, the presumption operative here—that grades are an objective measure of scholarly accomplishment—is an ideal, not a settled fact, but that, and the considerable difficulty with which bias and subjectivity are overcome, takes nothing away from the worthy importance of grading.

The results of the article are for the most part negative: It is shown that a prevalent method is inconsistent. No specific recommendation grows naturally out of this, if it is not that educators should cease to use and calculate GPAs. What should or could replace the GPA as an index of scholarly merit and accomplishment? If what is wanted is a single index that conforms to dominance where it exists and also fixes incomparable comparisons in some uniform way, the answer is that so long as grading is ordinal measurement, it is virtually certain that no single index can do this. It has been known at least since 1951, when Kenneth Arrow published his famous impossibility result (Arrow, 1951), that the amalgamation of a collection of strict simple orders (transitive, asymmetric, connected) into a collective strict simple order satisfying quite weak structural constraints (all of which apply obviously to the case of course and class ranks) is impossible. So that question is settled.

Grades and their averages have assumed an enormous importance in the United States today. After the wealth and class of one's parents, the sort of education that one receives is perhaps the most significant factor in determining one's social and economic class, and grade averages, from middle school through admission to graduate and professional school, are a primary instrument of selection and tracking. What is said and proven above shows that the current system of grading and averaging cannot support the great social and economic weight that burdens it. What is to be done? Here are a few modest proposals.

First and simplest, those who cannot counter the above arguments should ignore GPAs. That is pretty easy to do, since they are usually accompanied with transcripts or other records, and when they are not, policy can always require these.

Second, it is time that the great obscurities and inequities in grading and averaging be considered in a general academic and political setting. Some responsible authority should convene a committee of methodologically knowledgeable and concerned people to consider at least the following obvious questions:

What sort of uniform code to guide grading practices in secondary schools and higher education would eliminate, or at least hinder, shortcomings such as grade inflation, subjectivity, and variation among institutions, of the present congeries of systems?

What sort of uniform code for summarizing the information in a transcript or course record can strike a good compromise between too much information to handle and too little information to support academic decisions?

To what extent should educators at different levels be responsible to provide summary evaluations of students for the use of others: secondary school teachers for colleges and universities, college and university teachers for employers and graduate and professional schools?

These are difficult questions, both technically and politically. They need to be faced and answered. Our current practices yield neither justice nor truth.

APPENDIX: THE COMPARABILITY AND INCOMPARABILITY THEOREMS

We recall some definitions. A (student) *profile*

$$p = \langle p_1 \ldots p_i \ldots p_k \rangle$$

is an ordered k-tuple of grades from the set {E, D, C, B, A}. To simplify the argument we restrict consideration to profiles of the same number k of grades. Profiles p and q are *equivalent* if they have the same numbers of all grades. A profile p *strictly dominates* a profile q if for some permutations p' and q' of p and q, $p'_j > q'_j$ for some j, and $p'_i \geq q'_i$ for all i. p *simply dominates* q if $p'_i \geq q'_i$ for all i. Profiles p and q are *comparable* if either simply dominates the other, otherwise *incomparable*.

An *ordinal* scale is an assignment of numbers to grades that preserves the relation of *better than*.

The proof of the comparability theorem is simple and straightforward:

Comparability theorem. If p and q are comparable profiles of length k and γ is any ordinal scale, then

$$\text{p strictly dominates q iff } (1/k)\Sigma_i\, \gamma(p_i) > (1/k)\Sigma_i \gamma(q_i)$$

$$\text{p simply dominates q iff } (1/k)\Sigma_i\, \gamma(p_i) \geq (1/k)\Sigma_i \gamma(q_i)$$

$$\text{p is equivalent to q iff } (1/k)\Sigma_i\, \gamma(p_i) = (1/k)\Sigma_i \gamma(q_i)$$

Proof: Assume p and q to be comparable and γ to be any ordinal scale. Then p strictly dominates q iff p is nowhere worse and somewhere better than q, so

p strictly dominates q iff $\gamma(p_i) \geq \gamma(q_i)$ for all i, and for some j $\gamma(p_j) > \gamma(q_j)$
p strictly dominates q iff $\Sigma_i\, \gamma(p_i) > \Sigma_i\, \gamma(q_i)$
p strictly dominates q iff $(1/k)\Sigma_i\, \gamma(p_i) > (1/k)\Sigma_i\, \gamma(q_i)$.

The second and third clauses are proved analogously.

The proof of the incomparability theorem is less direct. Recall that it states that if p and q are incomparable profiles (in every permutation each is somewhere better than the other) then the order of their averages varies with scale: Change of scale can make their averages equal or either greater than the other. First a few definitions and conventions.

To simplify notation we denote the grades E, D, C, B, A by their values in the four-point scale. So, 0 represents E, 1, represents D, 2 represents C, 3 represents B, and 4 represents A. Thinking of grades in this way, for each grade $x \in \{0, 1, 2, 3, 4\}$ and each profile p, let *n(p,x)* be the number of grades in p equal to x. Then for each ordinal scale γ and each profile p,

$$\Sigma_i\ \gamma(p_i) = \Sigma_x\ n(p,x)\gamma(x).$$

A *monotone transformation* of numbers is a function that always assigns greater numbers to greater numbers and equal numbers to equal numbers. Clearly, every monotone transformation of any ordinal scale is an ordinal scale, and every ordinal scale is a monotone transformation of the coding in 0, 1, 2, 3, 4. In view of this, we simply identify ordinal scales with monotone transformations of {0, 1, 2, 3, 4}.

The *canonical permutation* of a profile p is that permutation of p in which the elements of p appear in increasing order. So if p is canonical and $i \le j$, $p_i \le p_j$.

We can now state and prove a lemma that immediately entails the Incomparability theorem:

Lemma. If q does not simply dominate p then for some ordinal assignments γ and δ

$$(i)\ \Sigma_x\ n(p,x)\gamma(x) = \Sigma_x\ n(q,x)\gamma(x)$$
$$(ii)\ \Sigma_x\ n(p,x)\delta(x) > \Sigma_x\ n(q,x)\delta(x)$$

Proof. Assume without loss of generality that p and q are both canonical and that q does not simply dominate p. Then for some index i, $p_i > q_i$. Hence all of p_i, $p_{i+1}, \ldots p_k$ are greater than or equal to p_i, while $q_i < p_i$, so p includes fewer grades lower than p_i than does q. From this it follows that for some grade y greater than 0 (i.e., better than E), there are more grades at least as good as y in p than there are in q, i.e.,

$$\Sigma_{x\ge y}\ n(p,x) > \Sigma_{x\ge y}\ n(q,x).$$

(Recall that p and q are ordered by $\le$.) Let g be the least such grade, and define the monotone transformation φ_g

$$If\ x < g\ then\ \varphi_g(x) = x$$
$$If\ x \ge g\ then\ \varphi_g(x) = x + m$$

where m is

$$\frac{\Sigma_x\ x[n(p,x) - n(q,x)]}{\Sigma_{x\ge g}\ [n(q,x) - n(p,x)]}.$$

Since $\Sigma_{x\ge g}n(p,x) > \Sigma_{x\ge g}n(q,x)$, the denominator of m is negative, not zero. Hence m is negative, positive, or zero accordingly as $\Sigma_i p_i$ is greater than, less than, or equal to $\Sigma_i q_i$.

We have:

$$\Sigma_x\ x[n(p,x) - n(q,x)] = \Sigma_{x\geq g}\ m[n(q,x) - n(p,x)] \qquad \Sigma_x\ n(p,x)x + \Sigma_{x\geq g}$$
$$n(p,x)m = \Sigma_x\ n(q,x)x + \Sigma_{x\geq g}\ n(q,x)m \qquad \Sigma_{x<g}\ n(p,x)\phi_g(x) + \Sigma_{x\geq g}\ n(p,x)\phi_g(x) = \Sigma_{x<g}$$
$$n(q,x)\phi_g(x) + \Sigma_{x\geq g}\ n(q,x)\phi_g(x) \qquad \Sigma_x\ n(p,x)\phi_g(x) = \Sigma_g n(q,x)\phi_g(x).$$

This establishes (i). To prove (ii) set $\delta = \varphi_g + 1$. This completes the proof of the lemma. The incomparability theorem now follows immediately:

Incomparability Theorem. If p and q are incomparable then there are ordinal scales in which p has a greater average than q, in which q has a greater average than p, and in which their averages are equal.

Notice that in the proof of the lemma we actually establish something a bit stronger, namely that:

If p, q, are any profiles, then p strictly dominates q iff $p'_i > q'_i$ for some i, and for all j, $p'_j \geq q'_j$, where p' and q' are the canonical permutations of p and q.

This is a consequence of the symmetry of all courses, that no grade is more important than any other, and it shows that canonical permutations live up to their name: Comparability is decided by the relations between canonical permutations.

NOTES

1. Thanks to David Beard, head counselor of Woodrow Wilson High School in Long Beach, California, to Carol Entler, Registrar of Scripps College, and to Edris Stuebner, erstwhile Registrar of the Claremont Graduate University, for generous help and information. Thanks to Kerry Odell of Scripps College for helpful comments, to an anonymous referee for suggesting the Incomparability Theorem, to the editor of this journal for patient counsel, and thanks above all to my colleagues Dale Berger, David Drew, and Charles Young for valuable counsel and criticism.
2. See Milton, Pollio, and Eison, 1986, chapter 2 and pp. 218–223, for a critical discussion of applications and interpretations of GPAs. The authors hold the GPA to be meaningless and recommend that it be abolished, but they do not mention the inconsistency described here.
3. Or did require this in 1991. See Siegel and Anderson, 1991.
4. Suppes and Zinnes, 1963, is a thorough development of the subject.
5. Masses and distances are squared in the Newtonian law of universal gravitation, but the presence of a constant specifying the units of mass and distance assures that the law is invariant for shifts of scale. Length squared, of course, gives the area of a square.
6. The simple truth is that all empirically adequate scales for the measure of weight and length are related by ratio or multiplicative transformations: weight in pounds = 2.2(weight in kilos). Empirically adequate temperature scales, including the Celsius and Fahrenheit scales, are related by affine or linear transformations: $F = 1/5C + 32$; $C = 5/9F + 160/9$. These permit averaging, but give no meaning to ratios. Absolute temperature scales, such as the Kelvin scale, which figure prominently in ideal gas laws, are another story, and this is not the place to tell it.

7. The Incomparability Theorem was suggested by a remark of a referee.
8. The puzzle was raised in a question by Dale Berger.
9. Young, 1990. The example is mine. There is no evidence that the author succumbs to the fallacy illustrated here. The inconsistency in grade averaging is however not mentioned.
10. That Maurice can have higher percentages in both subsets and a lower percentage overall is an instance of Simpson's paradox. See Simpson, 1951.

REFERENCES

Arrow, Kenneth J. (1951). *Social Choice and Individual Welfare. Cowles Commission Monograph* 12. New York: John Wiley and Sons.

Lang, David M. (1997). Accurate and Fair Class Ranks: One Step Closer with the Class Rank Index, *ERS Spectrum* 15(3) (Summer 1997): 26–29.

Milton, Ohmer, Pollio, Howard R., and Eison, James A. (1986). *Making Sense of College Grades: Why the Grading System Does Not Work and What Can be Done About It*. Foreword by Laura Bornholdt. San Francisco and London: Jossey-Bass Publishers.

Siegel, Judith, and Anderson, Carolyn S. (1991). Considerations in Calculating High School GPA and Rank-in-Class. *NASSP Bulletin* 75(537) (October 1991): 96–109.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society series B* 13: 238–241.

Suppes, Patrick, and Zinnes, Joseph L. (1963). Basic measurement theory. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter (eds.), *Handbook of Mathematical Psychology*, vol. I. New York and London, John Wiley and Sons.

Young, John W. (1990). Are validity coefficients understated due to correctable defects in the GPA? *Research in Higher Education* 31(4): 319–325.

Received August 1, 1998.