2007

The Injustice of Reducing the Number of Levels in a Grading System

William K.S. Wang
UC Hastings College of the Law, wangw@uchastings.edu

Author: William K.S. Wang

Source: Journal of Legal Education

Citation: 57 J. Legal Educ. 423 (2007).

Title: *The Injustice of Reducing the Number of Levels in a Grading System*

The Injustice of Reducing the Number of Levels in a Grading System

William K.S. Wang

Introduction

An issue faced by many law schools and administrators of standardized tests (e.g., the Law School Admission Test) is whether to move from a grading system with more levels to one with fewer levels. I examine the effect of such a move on the overall "fairness" or "justice" of grades or scores.

A system with only a few possible grades (such as Honors, Pass, Low Pass, Fail) may have certain advantages, such as avoiding the false aura of precision conveyed by a system with an extremely large number of possible grades. In addition, fewer possible grades may lead to a less competitive atmosphere.

Nevertheless, as long as some rational basis exists for assigning different grades in a given system, a decrease in the number of grade levels may result in less justice. To demonstrate this point, I hypothesize two test-takers, each with scores of 82 on a scale of 50-100. After formulating two different measures of unfairness, I show that unfairness may increase when converting from a scale of 50-100 to one of 5-10—which in effect rounds down the two 82 scores to 8, the equivalent of 80 on a 50-100 scale.

Two Possible Unfairness Indexes

Assume two individuals, L and H, take a test with a sufficient number of questions to generate rationally a 50-100 grading scale. Both test-takers receive scores of 82. The true scores, however, are 84 for L and 80 for H. In other

William K. S. Wang is a professor of law, University of California, Hastings College of Law.

I gratefully acknowledge the helpful assistance of my daughter, Karen Wang, and my wife, Kwan Wang, and of Professors Lisa Bingham, David Faigman, Heather Field, Ethan Leib, Aaron Rappaport, and Mark Tushnet. I especially appreciate the insightful advice of Dr. Peter Pashley, on the staff of the Law School Admission Council.

Journal of Legal Education, Volume 57, Number 3 (September 2007)

words, L's obtained test score is two points too low, and H's obtained test score is two points too high.[1]

Unfairness clearly exists. But how should it be measured? Many possible measures exist. For illustration, I describe two alternative unfairness indexes: the Simple Unfairness Index and the Squared Unfairness Index. The Simple Unfairness Index would *average* the absolute values of all errors (error defined as the difference between the obtained score and the true score). In this example, the absolute value of each error would be 2, and the Simple Unfairness Index would be 2.

A possible problem exists with the Simple Unfairness Index, however. People may feel that unfairness increases more than proportionately as the obtained score becomes less accurate. A grade twelve points off is more than twelve times as unfair as a grade one point off. If so, the Simple Unfairness Index does not account well for people's sense of injustice. A better unfairness index would increase *more than proportionately* as obtained scores became less accurate.

One such index would be the following: (1) for each obtained grade, calculate the arithmetic difference between the obtained grade and the true grade, (2) square that difference, and (3) take the average of all the squared amounts. In other words, average the *squares* of the errors. This unfairness index will be called the Squared Unfairness Index.[2] In this example of students L and H,

1. Grades could be inaccurate for two different reasons: (1) sub-optimal test questions or (2) sub-optimal grading (in essay questions, as opposed to multiple-choice).

To address the problem of suboptimal test questions, administrators could in theory implement a system in which teams of experts drafted and redrafted questions, with even more redrafting after pre-testing of the questions. Tests using these questions would generate more accurate grades.

In practice, grading essay examinations is an imprecise activity. A single grader reading the answers at different times might assign different grades, and different graders might assign varying grades as well. Nevertheless, the common assumption is that a central tendency exists and that the various grades are estimates of the accurate or true grade. Hypothetically, one might approximate the accurate or true grade by averaging all the different grades awarded by numerous graders each reading the test answers many times. Alternatively, one might try to approximate accurate or true grades through other hypothetical processes. Regardless of how one conceives the process, however, I assume in this article that a hypothetical "accurate" or "true" grade exists.

2. This Squared Unfairness Index is somewhat similar to the concept of "variance" in statistics. Variance is the average of the squares of the deviations from the *mean*. For discussion of "variance," see Wilbur G. Lewellen, The Cost of Capital ch. 2 (1969), *excerpted in* William W. Bratton, Corporate Finance: Cases and Materials 69-70 (6th ed., New York, 2008).

Were the disproportionate increase in unfairness less dramatic, one might raise the (absolute values of the) errors to a power between one and two before calculating the average to determine the unfairness index score. Were the disproportionate increase in unfairness truly dramatic, the appropriate unfairness index might average the cubes of the (absolute values of the) errors, rather than the squares. One might even raise the errors to powers greater than three before calculating the average.

For ease of exposition, I chose the Squared Unfairness Index. As long as the errors are raised to some power greater than one, however, the unfairness index will increase the Squared Unfairness Index would be 4: 2 squared + 2 squared (or 4 + 4), divided by 2.

Rounding from Two to One Significant Digit

As before, assume that two individuals, L and H, take a test with a sufficient number of questions to generate rationally a 50-100 grading scale. Both test-takers receive obtained scores of 82. The true scores, however, are 84 for L and 80 for H. Again, L's obtained test score is two points too low, and H's obtained test score is two points too high. Table I below illustrates this hypothetical.

TABLE I

	Test-taker L	Test-taker H
Obtained Score	82	82
True Score	84	80
Absolute Value of Error	2	2
Squared Error	4	4
Simple Unfairness Index (average of absolute values of errors)		2
Squared Unfairness Index (average of squares of errors)		4

Now, assume retention of the 50-100 scale but conversion of all scores to ones ending in zero by rounding to the nearest ten (the equivalent of a 5-10 scale). In others words, a score of 82 would be rounded down to 80 (or 8 on a 5-10 scale). The following data result:

TABLE II

	Test-taker L	Test-taker H
Obtained Score	80	80
True Score	84	80
Absolute Value of Error	4	0
Squared Error	16	0
Simple Unfairness Index (average of absolute values of errors)		2
Squared Unfairness Index (average of squares of errors)		8

disproportionately and should behave roughly like the Squared Unfairness Index in my examples.

With L, rounding down from 82 to 80 decreased accuracy by two points. With H, rounding down from 82 to 80 increased accuracy by two points. The Simple Unfairness Index remained the same. The Squared Unfairness Index, however, increased significantly because this index increases disproportionately as L's obtained test score becomes increasingly inaccurate.

Conclusion

This hypothetical suggests that if unfairness does not increase disproportionately as a test score's inaccuracy increases, reducing the number of levels in a grading system does not necessarily decrease fairness. On the other hand, if unfairness does increase disproportionately with inaccuracy, converting to a grading system with fewer levels probably decreases fairness.

If one accepts the second concept of unfairness, and if the principal goal is to minimize unfairness so defined, any test should be graded on a scale with as many levels as the examination can rationally generate. A law school should permit its faculty to grade on a scale with as many levels as the faculty can rationally handle. If a faculty member cannot or will not grade so precisely, he or she can always use a fraction of the grades available (for example, only numbers evenly divisible by two, three, ten, etc.). In other words, a numerical grading system with a large number of grading levels gives faculty members the option of choosing fewer levels, while a system with fewer levels does *not* give faculty members the alternative of choosing *more* levels.