Testing Unfairness - Mathematics and Simulations

G. Donald Allen
Department of Mathematics
Texas A&M University
College Station, TX 77843
gdonaldallen@gmail.com

Dianne Goldsby
Department of Teaching, Learning and Culture
Texas A&M University
College Station, TX 77843
dgoldsby@tamu.edu

Sandra Nite
Aggie STEM, Department of Mathematics
Texas A&M University
College Station, TX 77843
snite@math.tamu.edu

Abstract: This paper examines some novel issues of testing unfairness from perspectives of the student, teacher, but mostly the test itself. The literature usually examines testing unfairness between various groups, considering mainly group demographics. We consider cognitively differentiated groups and test difficulty to examine how groups perform on a particular test, both having defined distributions of ability and difficulty, respectively. Even controlling the statistical parameters of the test and/or the group, there can be vast differences in outcomes, i.e. exam averages. This holds even for high stakes testing, leading to the reality grades are almost always "adjusted" to conform to testing expectations, i.e. we curve exam scores. In this section, we consider yet another aspect of unfairness, and that is with the numbers embedded in the test itself, and how it affects class grades with respect to class parameters. The two principle parameters about the student are the average ability and the variation of those abilities. For the test, the two principle parameters are the average item difficulty and its variation. However, we need also consider the actual distributions of the student abilities and the test difficulties. These must be co-mingled in an intuitive manner. Various types of distributions of these populations and difficulties are considered. In addition, we need to model the actual densities of student abilities and item difficulties given the numbers of students and test items. For tests given to a relatively small number of students, there are other considerations such as the need to use simulations. We discuss the sensitivity of grades to parameters, distribution of grades, the effect of a small number of test takers, and the relative variation in standardized tests.

Keywords: Testing, item difficulty, class ability, normal distributions, beta distributions

1. Introduction

We constantly hear testing is unfair. Is it the testing, the tests or a combination of both? Is it the teacher or the student or the administration? Check "yes" for all. Indeed, students create unfairness for themselves in a variety of ways. We begin with a brief table of sources and review some. In the methods section, we illustrate that random effects do create surprisingly large effects and that in our continuing efforts to make tests fairer, we may be simply fooled by randomness (Taleb, 2001). A brief table of sources reveals a mix of sources, many of which only the teachers knows and can ever know.

For example, many teachers, knowingly or not, telegraph to their students a spectrum of information about the test. For multiple choice tests, the list of why's is endless, but there are other unfair practices that widen the gap between good and poor students. In some way, the practice of testing for almost every student type is unfair by its very nature. To frame our intent at examining more interesting and novel unfairness issues, we review some of the more standard issues of unfairness and limitation of multiple-choice tests. A multiple-choice question is generally "defined as a question in which students are asked to select one alternative from a given set of alternatives in response to a question stem" (Torres, Lopes, Babo, &Azevedo, 2011, p. 1).

The purpose of the distractors is to appear as plausible solutions to the problem for those students who have not achieved the objective being measured by the test item. Conversely, the distractors must appear as implausible solutions for those students who have achieved the objective. Only the answer should appear plausible to these students. (Burton, Sudweeks, Merrill, & Wood, 1991, p. 3).

The issues include question misinterpretation, the evaluation of knowledge beyond the range of options provided, difficulty in phrasing for identical interpretation by all students, the encouragement of guessing,

anxiety, and errant test-taking strategies. There are certain learning outcomes that are difficult to measure with this kind of test items (Torres et al., 2011) and students' scores may not be fairly representative of true achievement unless the scores are transformed in some way to reduce the adverse effects of guessing (Downing, 2003).

Testing Unfairness - Sources			
Student	**Teacher**	**Test**	**Administrative**
Cognitive	Style	Number of questions	Practice time
Study practice	Curriculum	Difficulty distribution	Curriculum
Group effects	Rubric rigor	Language	Stress
Distribution of abilities	Telegraph	Answer traps	
Response to grade	Prep time	Random effects	
Testing strategies	Halo effect	Number of questions	
Anxiety	Time Management	Question structure	
Random effects		Question type (MC, Essay, etc)	

Table 1 - Sources of Testing Unfairness

The topic of testing unfairness is vast, mostly unexplored, but ever present in our testing of students. Each of the sub items of Table 1 has its own story to tell. In this paper, we consider only the random effects of testing. This involves the statistics of the classroom, insofar as their abilities, and the statistics of the test. So we look at administering the same test year-after-year, with only the class changes. Of particular interest will be just how much the class average can range. We also study the effects of how many students take the test. The smaller the numbers, the greater possibilities for vastly different results. So, we will introduce a theory involving student abilities and item difficulties. They will be cast in the form of probability distributions. From this we obtain a "theoretical" grade. Unrealistically, this grade would be the class grade for an infinite sized class with an infinite number of items. To obtain realistic results, we will simulate the grade computation by taking finite samples of the appropriate distributions, and then take the test for each student in the sample.

2. Basic Assumptions

To arrange the basic parameters, we make the some definitions. Following the tenets of the scoring correction literature, we assume that all student abilities are calibrated to be in the range $[0,1]$, with the ablest students having ability 1 and the least able having ability 0. In reverse item difficulties are also in the range $[0,1]$ with the most difficult problems at 0 and the easiest problems at 1. This inverse scale is in observation to standard item analysis (Barnard, 1995). For student with ability p attempting to solve a problem of difficulty q the probability of the student solving the problem is assigned to be

$$P(p \mid q) \begin{cases} 0 & \text{if } p < q \\ \frac{1}{2} & \text{if } p = q \\ 1 & \text{if } p > q \end{cases}$$

This means simply if the student knows the material at the level of difficulty q then the correct answer is given. If otherwise, the incorrect answer is given. The middle case, with $p = 1 - q$, is considered a "draw" with the resulting probability assigned as 1/2. In all of our calculations, the middle case is vanishingly rare, essentially non-existent. When a test is comprised of one problem with one correct answer, there is little here. However, in the following we consider a number of students and a number problem difficulties, and then compute the class average. Before the details, we give the fundamental tenets of scoring correction (Budescu, Bar-Hiller, 1993; Hales and Marshall, 1972, Rowley and Traub, 1977, and Frary, 1988).

1. The person tested either has the knowledge to answer it correctly or doesn't.
2. With the knowledge the person will answer correctly; without the knowledge, the person will guess.
3. Each incorrect response was the result of a random guess (among all options given).

These must be treated in turn. The first is on knowledge and difficulty, and item 1. Guessing is postponed but is rather easy to handle within our general framework. Essentially, either a student knows the correct answer or not Let us compute the **true grade** for a student of ability p taking a test of difficulties $q_i, i = 1..n_t$. By the

formula above it will be

$$T = \frac{1}{n_t} \sum_{i} \sum_{p \geq q_i} 1$$

This is simply the proportion of the number of test items n_t where the difficulties $1 - q$ are less than the ability p. Multiply by 100 to get the percentage. We use these terms, percentages, and proportion, interchangeably. Now assume a class of n_s students with abilities given by $p_j, j = 1..n_s$. The true grade for each student follows the formula above with the appropriate p_j replacing p. The **true class average** , therefore, is given by

$$T_C = \frac{1}{n_s n_t} \sum_{j} \sum_{i} \sum_{p_j \geq q_i} 1$$

So far, this amounts counting and then averaging. The word *true* will be clarified a bit later. Note, we continue to omit the special case when the student ability is p and the problem difficulty is $1 - p$. This is inconsequential. One reason is that we will use random number generators that produce twelve digit numbers for both probabilities, and the other is the matter of fuzziness. Namely, it can never be known exactly what the ability of either a student is or what the item difficulty is. This is reasonable. It means we can omit the possibility. However, this fuzziness actually constitutes one of the several simulation problems.
First, it is important to give a form to the student (ability) and testing (difficulty) parameters in the terms of their probability densities. We consider two forms of such densities, the normal and the beta. It is reasonable to envision either of these forms. Even prior to this we need to generalize the relations to infinite distributions for both student abilities and item difficulties. For example, suppose we are presented with a large number of item test with a huge number of students. We need to express the formulae above in more generality. These will be used later.

Continuous Densities. Of course, not all students have the same ability, and not all test items have the same difficulty. We need to assume both have a probability density. Define μ_s to be the mean or average ability of the student population with standard deviation σ_s. We write $f_s(p)$ to be the probability density of student abilities. Similarly, let $f_d(q)$ be the probability density of item difficulties having mean μ_d and standard deviation σ_d. Note $0 \leq p, q \leq 1$. Some consideration should be given about how the difficulty of an item is measured.
The generalization of the true average class exam score in these circumstances is

$$T_C = \int_0^1 f_a(p) \int_{1-p}^1 f_d(q) dq\, dp$$

This applies for discrete densities, as above, in the cases when the distributions are point masses and $\int_0^1 \delta_p(t) \int_{1-t}^1 \delta_{1-p}(s) ds\, dt = \frac{1}{2}$, where $\delta_p(x)$ is the Dirac density supported at p. This is the case when the integrals are defined in the usual symmetric manner as limits of step functions. This number is in $[0,1]$, and therefore to get the grade percentage multiply by 100. Thus, the more general formula reduces to the previous one with discrete densities.
With continuous densities as we use below in computations, there is assumed a test with both an infinite number of questions and an infinite number of students. Yet in practice, T_C is near to the simulated average, when numbers of questions and students are reasonably large. We state the interesting result consistent with these ideas.
Theorem 1. If $f_a(x)$ and $f_d(x)$ are continuous symmetric densities about their means μ and $1 - \mu$ respectively, with $0 < \mu < 1$ then

$$\int_0^1 f_a(p) \int_{1-p}^1 f_d(q) dq\, dp = \frac{1}{2}$$

Here, symmetric means $f(\mu + \varepsilon) = f(\mu - \epsilon)$ for all appropriate values ϵ. The proof is not difficult but takes us beyond the scope of this paper. For our purposes, it is best to regard the densities as pertaining to those normal, or even uniform. But the grade T_C follows from the theorem. When the means of difficulty and ability are in this relation, μ and $1 - \mu$, they are said to be **complementary**.
We will consider the types of densities (or distributions) used in the simulations. They are the normal and beta densities. Both exhibit properties we expect and see in student populations and test item difficulties. For

example, on a test, it is common for there to be just a couple of challenging problems, with most of the others relatively routine. It is also common to see abilities of the class to be somewhat biased toward the higher end, with fewer students of only slight ability. The normal and beta densities have these properties.

3. Normal and Beta Distributions

The familiar normal density has the form

$$N(\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty$$

Its graph shows the traditional bell-shaped curve. In this example, we consider student abilities to be normally distributed with mean μ_1 and standard deviation σ_1. The range of item difficulties is also assumed to be normally distributed with mean μ_2 and standard deviation σ_2. We take the standard deviations to keep the distribution substantially within $[0,1]$. This computes various probabilities for the bi-normal distribution arising in testing. For example, with

μ_1 = 0.5	σ_1 = min(μ_1, 1 − μ_1)
μ_2 = 0.6	σ_2 = min(μ_2, 1 − μ_2)

The true class average, for these means and standard deviations, is

$$T_C=\frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty}e^{-\frac{(t-\mu_1)^2}{2\sigma_1^2}}\left(\int_{1-t}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(s-\mu_2)^2}{2\sigma_2^2}}ds\right)dt=0.680$$

We can apply a change of variables to both variables to transform this to

$$T_C=\frac{1}{\pi}\int_{-\infty}^{\infty}e^{-w^2}\int_{\frac{1-\sqrt{2}\sigma_1 w-(\mu_1+\mu_2)}{\sqrt{2}\sigma_2}}^{\infty}e^{-v^2}\,dvdw$$

The true grade is given by

$$\frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{\infty}e^{-\frac{(t-\mu_1)^2}{2\sigma_1^2}}\left(\int_{1-t}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(s-\mu_2)^2}{2\sigma_2^2}}ds\right)dt$$

We can apply a change of variables to both variables to transform this to

$$T_C=\frac{1}{\pi}\int_{-\infty}^{\infty}e^{-w^2}\int_{\frac{1-\sqrt{2}\sigma_1 w-(\mu_1+\mu_2)}{\sqrt{2}\sigma_2}}^{\infty}e^{-v^2}\,dvdw$$

It is easy to see, by looking at respective ranges of integration that if μ_1 or μ_2 increase, then T_C increases. Similarly, if μ_1 or μ_2 decrease, then T_C decreases. This is as expected: make the students better or the test easier, and grades go up, and vice versa. Observe as well the true grade is an invariant of the sum $\mu_1+\mu_2$.In the special case where the sum is one, i.e. $\mu_1=1-\mu_2$, the true grade formula becomes

$$T_C=\frac{1}{\pi}\int_{-\infty}^{\infty}e^{-w^2}\int_{-\frac{\sigma_1 w}{\sigma_2}}^{\infty}e^{-v^2}\,dvdw$$

Thus, the true grade is independent of the means. Moreover, if we consider just the integral

$$T_C=\frac{1}{\pi}\int_{-\infty}^{\infty}e^{-w^2}\int_{cw}^{\infty}e^{-v^2}\,dvdw$$

it is a simple matter to see it is invariant under c, which is to say, also invariant under changes in the respective standard deviations. In the case, $c=1$, this evaluates to $\frac{1}{2}$, thus substantiating our theorem.

The standard deviations were taken to keep the distributions more-or-less in $[0,1]$, though this is not really necessary. As you see, because this test is relatively easy, with $\mu_2=0.6$, the test favors higher grades even though the mean ability for all students is $\mu_1=0.5$. Simply by looking at the expression for T_C we see the relations between the four variables is nonlinear.

Figure 1 shows what a typical theoretical grade vs. difficulty looks like when using normal densities. In this graph $\mu_1 = 0.5$ and $\sigma_1 = 0.083$. All such graphs have this general sigmoidal appearance as μ_1 ranges in $[0,1]$. The x-axis below is the difficulty mean μ_2 with $\sigma_2 = \mu_2/6$.

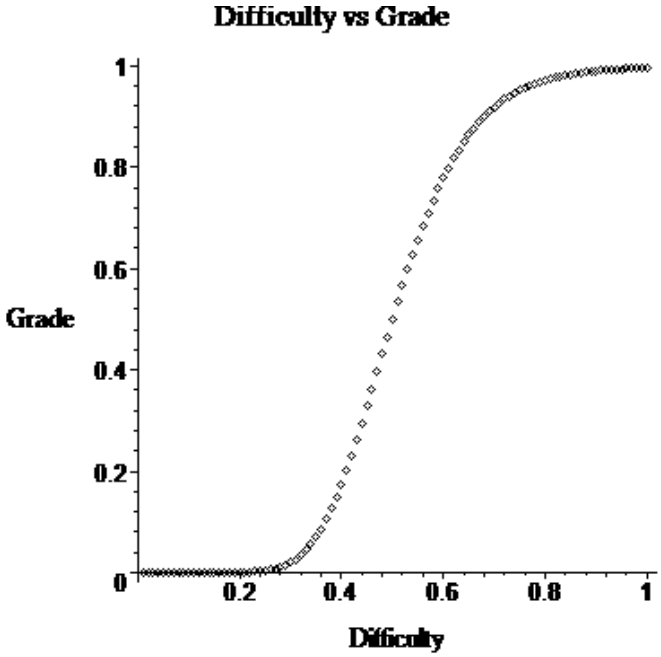

Figure 1

This suggests the expected grade $T(\mu_1, \sigma_1, \mu_2, \sigma_2)$, now expressed as a function of its four primary parameters, is an increasing function of μ_1, and in fact we note that $\frac{\partial T}{\partial \mu_1} > 0$. Moreover, it is relatively easy to show that $\frac{\partial T}{\partial \mu_1}$ attains its maximum when $\mu_1 = 1 - \mu_2$, that is to say

$$\left.\frac{\partial^2 T}{\partial \mu_1^2}\right|_{\mu_1 = 1-\mu_2} = 0$$

It is of some interest to note that even a small change in student ability can magnify the class grade outcome for such balanced or nearly balanced scenarios (i.e. $\mu_1 = 1 - \mu_2$). These results depend on the essential double symmetry of the normal distributions. The smaller σ_1 is, the larger is the magnification, and moreover, this occurs proportional to the square of σ_1. Typically, the magnification is about triple the change in the mean student ability. For example, a 1% change in mean class ability can change the class grade by 3% or more, while one would expect a corresponding 1% change. One upshot of this is that tests with a very narrow spread of problem difficulties or student abilities can have a dramatic effect on class grades. This is a sensitivity aspect of testing of grades in their correspondence to student ability or problem difficulty. One way to minimize the magnification is to administer a test with a wide variety of difficulties. Finally, note

$$T \to 0.5 \quad \text{if} \quad \mu_1 \to 1 \text{ and } \mu_2$$
$$T \to 0.5 \quad \text{if} \quad \mu_1 \to 1 \text{ and } \mu_2$$
$$T \to 0.5 \quad \text{if} \quad \mu_1 \to 1 \text{ and } \mu_2$$

Beta Distributions

The beta density on [0,1] is defined by

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $\alpha, \beta > 0$, and where Γ denotes the gamma function (Johnson, et. al., 1995). The values α and β are

called the **shape** parameters of the beta density. When $\alpha, \beta \geq 2,$ it has the general shapes shown in Figure 2. The mean or expectation is given by

$$\mu = \frac{\alpha}{\alpha + \beta}$$

and standard deviation by

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

This density, which can be biased toward zero or one by selecting α and β appropriately, seems a better model for student abilities or item difficulties because they are first confined to $[0,1]$ but also they may be more representative of the actual situation. It is far easier to imagine for an average class the distribution of rather low abilities is lower than that for greater abilities. It is also easy to imagine for test item difficulties there are more easy problems that difficult ones.

This does seem to contradict the use of normal densities, as we are accustomed to with the IQ test. Yet it is important to remember IQ scores are normalized to $\mu = 100$ after the IQ test is taken. It may also account for at least college admission standards that favor better students. It would be expected that plotting student admissions would bias toward the upper end. This example shows the Beta density with the lower mean for $\alpha = 3$ and $\beta = 5,$ and the one with the greater mean Beta density for $\alpha = 7$ and $\beta = 2.$ It is always the case if $\alpha > \beta \geq 2,$ the distribution will be biased toward the right.

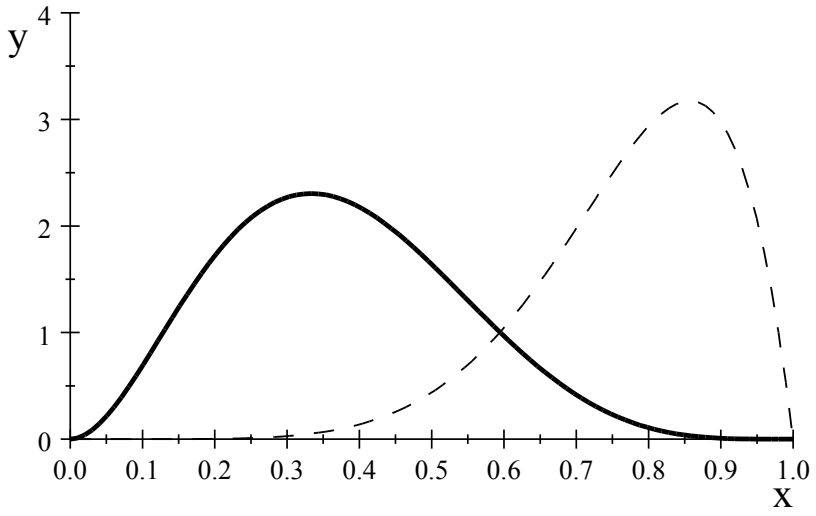

Figure 2. Beta densities

It is significant these are not symmetric. Suppose we look at two balanced profiles with means μ and $1-\mu.$ This implies we have $\mu = \frac{\alpha}{\alpha+\beta}$ and $1-\mu = 1-\frac{\alpha}{\alpha+\beta} = \frac{\beta}{\alpha+\beta}$. Thus, for each α and β, there is a pair of ability and difficulty profiles. We compute the expected true grade as

$$T = \int_0^1 f(t)\left(\int_{1-t}^1 g(s)ds\right)dt$$

where the "complementary" densities are given by $f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ and $g(x;\beta,\alpha) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\beta-1}(1-x)^{\alpha-1}$. It is a simple matter (change of variable and then change of order) to show that

$$T = \int_0^1 f(t)\left(\int_{1-t}^1 g(s)ds\right)dt = \int_0^1 g(t)\left(\int_t^1 g(s)ds\right)dt = \int_0^1 f(t)\left(\int_0^t g(s)ds\right)dt$$

Since the sum of the two right-hand integrals must be one, it follows that $\int_0^1 f(t)\left(\int_{1-t}^1 g(s)ds\right)dt = \frac{1}{2}$. Thus the true mean enjoys the same property as in the normal distribution case, even through symmetry is not present. A general theorem for the beta distributions can be proved.

Theorem 2. We have the four positive shape coefficients $\alpha_1,\ \beta_1, \alpha_2,$ and $\beta_2.$ with α_1 and β_1 referencing student ability, and with α_2 and β_2 referencing test difficulty. Then the true grade $T = \int_0^1 x^{\alpha_1-1}(1-x)^{\beta_1-1}\int_{1-x}^1 t^{\alpha_2-1}(1-t)^{\beta_2-1}dtdx$ satisfies the following

- Given $\alpha_1, \beta_1, \alpha_2$ are constant, then If β_2 increases, then T decreases and conversely.
- Given $\alpha_1, \beta_1, \beta_2$ are constant, then if α_2 increases, then T increases and conversely.
- Given $\beta_1, \alpha_2, \beta_2$ are constant, then if α_1 increases, then T increases and conversely.
- Given $\alpha_1, \alpha_2, \beta_2$ are constant, then if β_1 increases, then T decreases and conversely.

The proof, while not difficult, is a bit too long for inclusion here.

From these, it is easy to conclude related statements about the respective means and the true grade. If you prefer to set the mean and standard deviation for the Beta distribution, it is possible to determine the shape parameters by simply solving the system

$$\frac{\alpha}{\alpha+\beta} = \mu$$

$$\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = \sigma$$

for α and $\beta.$ Rewritten, there obtains an alternative form for $\sigma,$

$$\sigma = \mu\sqrt{\frac{1-\mu}{\alpha+\mu}}$$

It is also easy to see that

$$\frac{\alpha}{\beta} = \frac{\mu}{1-\mu}$$

Therefore, along lines with the form $\beta = c\alpha,$ the mean remains invariant. However, such an assignment does not leave the standard deviation invariant. Beginning with the mean and standard deviation, one solves to obtain the related shape parameters given by

$$\alpha = -\frac{1}{\sigma^2}(-\mu^2+\mu^3+\sigma^2\mu)$$

$$\beta = -\frac{(1-\mu)}{\mu\sigma^2}(-\mu^2+\mu^3+\sigma^2\mu)$$

Consider the normal case discuss above with the mean and standard deviation parameters as given. Let's use these to compute the simulated average grade for the respective beta densities. The respective shape parameters are $\alpha_1 = 4.0,\ \beta_1 = 4.0, \alpha_2 = 7.5,$ and $\beta_2 = 5.0.$ The average grade is given by

$$\int_0^1 x^{\alpha_1-1}(1-x)^{\beta_1-1}\int_{1-x}^1 t^{\alpha_2-1}(1-t)^{\beta_2-1}dtdx = 0.677$$

which is remarkably close the average when using the normal densities. Note, these integrals are generally not computable in closed form, that is by a formula, except of course when the shape parameters are integers. In these cases, the resulting formula is merely unwieldy. Just as with the normal distributions, they are computed numerically. The beta densities graphed are

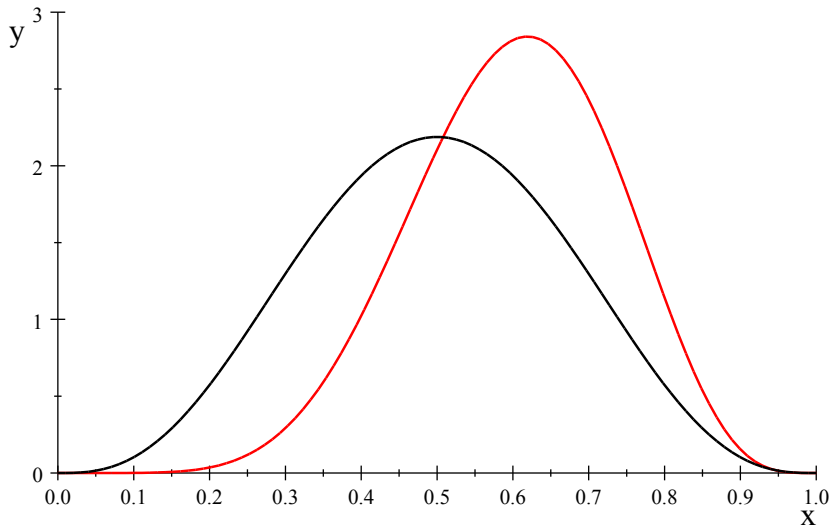

Figure 3 - Beta densities

Accommodating multiple choice tests is relatively easy in the framework set out. Indeed, adhering to the first principle, we will assume that if the student's ability p exceeds difficulty $1-p$,the students answers the question correctly. In alternative case the student, the student gets the wrong answer. However, when the answers are multiple choice, with say k, total distractors, the student guesses. And of those guessed at random, about $\frac{1}{k}$ of the time, the student guesses correctly. This gives the average class corrected grade of

$$T_{Ck} = \frac{1}{k} + \frac{k-1}{k} T_C$$

As is apparent, this raises the overall grade. Suppose a true-false exam $(k = 2)$ is given, where $T_C = 0.6.$ Then $T_{Ck} = \frac{1}{2} + \frac{1}{2} 0.6 = 0.8.$ Thus, when the instructor sees the class with an average of 80% on such an exam, it really means the demonstrated knowledge is but 60%. This calculation is reminiscent of scoring correction formulae, used in reverse. In our simulations, we will report the true and corrected grades based on $k = 5$ answer choices. Accommodating partial credit can be done similarly, but exactly how it is done is rather teacher dependent. For the SAT exam, there are strict grading rubrics applied that attempt to delimit individual preferences for partial credit.
Thus, the word "true" refers to the strict grade construction without guess and partial credit factors.

4. The Simulations

As noted earlier, the continuous normal and beta distributions are fine but basically assume an infinite population of students and an infinite number of test items. Therefore, they can't be used to simulate actual classroom performance, where there may be as few as a dozen students and perhaps only ten test items. This seems to be what happens from year to year as student populations change. We will see rather large spreads of the class grades even when the test remains the same year-by-year. Normally, we will administer (i.e. simulate) the tests over a 25 year period. This implication will be that there is really no such thing as a fair test. While standard deviation may be reasonably small, it is the range of averages that is more telling, mostly because it can happen.
As to the precision of student abilities and item difficulties, it is natural to expect these are subject to some variation. That is when determining how a student may do on a particular item, both the ability and the difficulty are unknown precisely. In fact, it is reasonable to argue they are only approximate. For example, if the student's ability is given as p it should be reasonable to expect

$$p = \overline{p} + E_p$$

where $\overline{p}$ is the mean ability and E_p is some stochastic term. For example, E_p could be a normal random variable with zero mean and small variance. However, simulations with this consideration included do not change outcomes to any significant extent. Thus, they are omitted.
In using normal distributions, it is certainly the case that the natural domain extends over the interval $(-\infty, \infty)$. One may surmise this can affect the results significantly, and therefore truncated normal distributions should be

used (Johnson, et. al. 1994). However, over the means and standard deviations we consider for reasonable classes and exams, outcomes, particularly grade ranges, change only very slightly. Indeed, we will usually set standard deviations of abilities and difficulties quite small, and still the grade ranges are remarkably large.

The algorithm. The description of this model is relatively simple to program for any with even modest programming skills. No details are omitted. First select *n* and *m,* the number of students and number of test items respectively. Now select a population *S* of *n* student abilities, s_i, $i = 1,...,n$, and a population of *D* of *m* test item difficulties, $d_i, i = 1,...,\mathrm{m}$. For student *i,* the score G_i is the number of occurrences where $s_i \geq 1 - d_j, j = 1,..,m$, and the grade is $T = 100\frac{G_i}{m}$. The class average is the average of the individual student grades. The populations are usually selected according to some rule, such as those discussed above.

As regards the condition where the ability is exactly complementary difficulty, and the conditional probability should be $\frac{1}{2}$, this does not affect outcomes either. Both abilities and difficulties are randomly selected according to the prescribed distributions have twelve digits, the upshot being this essentially never occurs, as checksums concur. What seems to be the problem, that is the deviations from true or expected grades is the granularity of the selected distributions when class sizes or test items are relatively small. It is easy to argue yearly that student abilities have just about the same distribution, as given by average and standard deviation. However, these two parameters are gross when compared with various possible samples having those parameters. This could explain why some teachers lament their class is not good this year or really good. It is because they are!

In our simulations, we compute $\overline{T}$, the average true grade over all the simulation years, and the corresponding standard deviation $\overline{S}$. The corrected true grade corresponding to $k \geq 2$ multiple choices will be $\frac{1}{k} + \frac{k-1}{k}\overline{T}$, with standard deviation $\frac{k-1}{k}\overline{S}$, as is easy to derive. So, multiple choice exams will enjoy a higher overall average and lower standard deviation, certainly making the test look better, maybe more fair.

Normal Distributions. In the first set of simulations, we use for item difficulty and student ability normal distributions such as those shown in Figure 4. In this example, the means are 0.6 (ability) and 0.5 (difficulty) with standard deviations 0.2 and 0.17 respectively. The theoretical mean, or true grade, is 77.88.

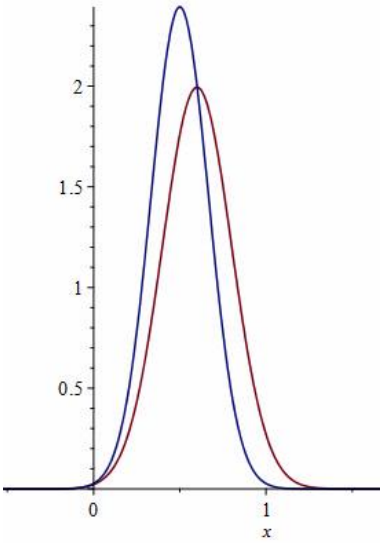

Figure 4 - Normal densities

We use these data in simulating a fixed test of 10 items given to 20 students. We perform the simulation over 25 cycles with new student populations with identical statistics taken each cycle with t. We have added a renormalization feature where for the typical student population is normalize to the precise mean. The class average with standard deviation is illustrated in Table 2.

Year	Grade	Standard Deviation	Year	Grade	Standard Deviation
1	88.00	19.89	14	70.50	24.17
2	78.00	28.58	15	83.50	18.72
3	71.00	30.42	16	80.50	20.12
4	84.00	20.88	17	87.50	19.16
5	79.50	25.64	18	68.50	28.70
6	79.50	17.91	19	84.00	22.34
7	88.00	16.73	20	71.50	27.77
8	73.50	23.23	21	88.00	12.40
9	72.00	30.02	22	79.00	28.82
10	78.50	21.59	23	75.00	27.24
11	83.00	19.22	24	64.00	29.81
12	78.00	28.21	25	75.50	23.28
13	73.50	18.43			

Table 2 - Variation in grades over 25 cycles

For this example, the average over all 25 cycles is 78.16, with a standard deviation of 6.63. The minimum average is= 64.00, and the maximum average is 88.00, giving a range of 24 points or two letter grades. The reason for this seriously large range is the low number of items and small class size. For the same 10 item test, now with a class size of 100, the minimum average is 76.00, the maximum average is 80.40 giving a range of 4.40 over a 25-year cycle. In all cases, there is a large standard deviation of grades for each cycle.

When the true grade is relatively high, such as in this example, the variation is less. We now consider tests and classes with the following parameters: Student: Average ability 0.4. Standard Deviation 0.0667; Test: Average Difficulty 0.6 Standard Deviation 0.1. Again, the distributions are normal. Results are displayed in Table 3. The theoretical mean, or true grade, is 50.00. We first consider various class sizes for a 33 item test.

Class size	Average Grade	Max Grade	Min Grade	Range
20	51.91	61.97	44.09	17.88
50	49.84	56.24	42.30	13.94
100	47.51	44.52	50.94	6.42
200	47.30	50.83	44.94	5.89
2000	47.47	66.04	45.61	1.86

Table 3 - Normal distributions and grades for class sizes

The average is, of course, the average of the averages over the 25 cycles. It varies according to the selected random samples for the test, for which we normalized the mean difficulty to be exactly 0.6, and the class abilities, which change from cycle to cycle with the prescribed mean and standard deviation. Thus, the class itself is a random sample, and as such it does have a natural mean and standard deviation, close to but not exactly the prescribed values (μ=0.4 and σ= 0.0667). Note, as the class size increases the spread of average grades decreases, but for even reasonably sized classes the range is more than a letter grade. It is no wonder we curve tests and constantly question how we are teaching. Finally, note that all these values themselves range quite a bit from simulation to simulation. Random effects are quite real and perhaps have a greater effect than one might imagine.

Beta Distributions. We consider the beta distributions with the shape parameters α_1=6.0 and β_1=4.5 for ability, together with and α_2=1.2, and β_2=0.9 for difficulty. For such distributions, the theoretical grade is 66.26. The respective means and standard deviations are μ_1=((6.0)/(6.0+4.5))= 0.571, σ_1= 0.146, μ_2=((1.2)/(1.2+0.9))= 0.571, and σ_2= 0.281. These have the general appearance as shown in Figure 5.

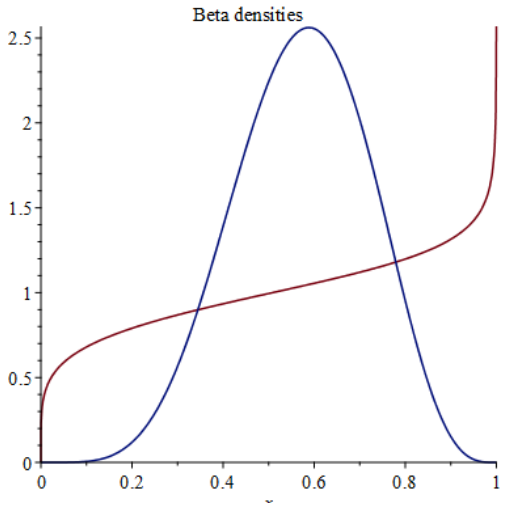

Figure 5 – Two beta densities

As you can see, the student abilities (blue) have a roughly normal shape where student abilities range around 0.6 but taper off at the ends, 0 and 1. The test, on the other hand, has a high proportion of easy problems (vertical asymptote), with the proportions of more difficult problems much smaller. This is not unlike tests we give, a sufficient number of easy problems to give the students a confidence boost, but many problems of graduated difficulty. Such shapes are not possible using normal distributions. To illustrate better, we illustrate this distribution with the approximate number of questions for a 33 item test with these shape parameters. (See Table 4.)

	Difficulty		
	From	To	Number of Questions
1	0.0	0.1	1.86
2	0.1	0.2	2.44
3	0.2	0.3	2.74
4	0.3	0.4	2.98
5	0.4	0.5	3.18
6	0.5	0.6	3.38
7	0.6	0.7	3.59
8	0.7	0.8	3.82
9	0.8	0.9	4.13
10	0.9	1.0	4.88

Number of questions computed as

$$33\left(\frac{\Gamma(1.2+0.9)}{\Gamma(1.2)\Gamma(0.9)}\right)\int_m^M x^{1.2-1}(1-x)^{0.9-1}dx$$

where the m and M are the From and To values in the table

Table 4 - Numbers of items by difficulty

This table reveals in numbers what the graph expresses, a number of very easy problems with problems of all other difficulties in significant proportions. For example, there are about two quite difficult problems. Very few, if any in this class will be successful at these items. On the other hand, there are nearly five very easy questions, which almost all students in the class will solve correctly. Next, we applied this test with 33 items and given to 40 students for twenty-five cycles. Results are shown in Table 5.

Year	Grade	Standard Deviation	Year	Grade	Standard Deviation
1	66.06	17.68	14	60.73	18.85
2	67.64	17.82	15	60.30	16.74
3	62.18	16.76	16	62.00	16.95
4	59.64	15.99	17	56.24	17.62
5	63.15	15.43	18	59.09	16.40
6	55.52	16.03	19	61.58	16.32
7	63.39	19.46	20	58.06	15.25
8	61.15	19.19	21	63.76	18.28
9	64.18	18.75	22	55.21	16.21
10	61.27	16.78	23	57.21	16.33
11	64.00	16.73	24	59.52	17.76
12	58.42	19.16	25	60.55	16.34

Table 5 - Variation in average grades for beta densities

However, for each cycle we modified the test by about 1% for each item, normally distributed. Over the 25 cycles, the overall average was 60.79, with standard deviation = 3.14. The minimum average was 55.21; the maximum average was 67.64. Interestingly, the range of grades was 12.42, implying that by essentially the same test to statistically the same class over 25 cycles, the range was more than one letter grade. Now we repeat the process for the same test for 25 cycles to various classes statistically identical. Typical results are shown in Table 6. As a general rule, if the test has fewer items, the ranges will become larger. This leads us to conclude that giving even the same test year after year can yield results of wide variation, even if the classes are statistically the same.

Class size	Average Grade	Max Grade	Min Grade	Range
20	66.69	72.12	60.30	11.82
50	68.04	73.46	62.97	10.49
100	67.69	69.69	65.27	4.42
200	66.74	68.65	64.83	3.82
2000	65.49	66.04	64.89	1.15

Table 6 - Beta distributions - class grades vs. class size

What is important to note is that with a 33 item exam, the test average for various sets of 25 cycles varies quite a lot. We selected a test to roughly match the theoretical average. Another point to note is the decrease in the range as the number of students increases. However, a class size of 2000 is extreme; it is more reminiscent of high stakes and placement exams. Finally, one may inquire if there is a significant difference in outcomes had we used normal distributions for the student abilities and item difficulties instead of the beta distributions. In fact, there are. For one thing, the shapes are radically different.

5. Conclusions

When we refer to the same class taking the same test, we always mean the same class statistically. That is, classes having the same distributions. This is the fundamental assumption under which schools operate every year. It allows for a fixed curriculum and relatively fixed end-of-course exams. It permits the educational system the stability it has. What our simulations do show is that for small item number tests and/or small classes, there arise wild variations in class averages. We note that changing the test, even by small amounts, can also cause wide swings The instructor who gives the same 10 item quiz every year in say Calculus I, should by our simulations experience quite different results from year to year. The department that gives common exams to all sections will see wide swings in section averages, for the two reasons that each section while statistically the same, is different and each section taught by a different instructor sees different things in the classroom. In both cases, the small quiz and the large common exam there are other unfairness issues. They are often with the teacher, in the forms of (1) how much information the teacher telegraphs to the student, or (2) how well the teacher teaches. In the common exam scenario, the exam makers must carefully pick questions and inform instructors of some sort of exam shell. Otherwise, results will and do vary widely, as experience has shown.

One of several problems in putting this to the general application from known data, is that almost all data comes from known tests for various student groupings. The four parameters we have used, two means and two standard deviations, plus the distribution forms are combined. This means it is necessary to separate them from the collective whole. This will be the topic of another paper.

Nonetheless, in summary we may say when it comes to testing, the multitude of unfairness forms is large, and how they intertwine is not well understood. The simulations have shown what can happen when we accommodate only student abilities and item difficulties, and here variations are considerable.

6. References

1.Barnard, J. J. (1995), Item Analysis in Test Construction pp. 195-206, in Geofferey N Masteres & John P Keeves, Advances in Measurement in Educational Research and Assessment, Pergamon.
2.Budescu, David and Maya Bar-Hillel, (1993),To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring Journal of Educational Measurement, Volume 30, Issue 4, pages 277--291.
3.Burgos, Albert, (2004) Guessing and Gambling", Economics Bulletin, Vol. 4, No. 4 pp. 1-10.
4.Burton,S., Sudweeks, R., Merrill, P., & Wood, B. (1991). How to prepare better multiple-

choice test items: Guidelines for university faculty. Retrieved from https://testing.byu.edu/handbooks/betteritems.pdf July 8, 2015
5.Hales, Loyde, and Marshall, Jon Clark, Essentials of Testing, Addison-Wesley, Reading, Mass.,1972.
6.Rowley and Traub (1977) Formula scoring, number right scoring, and test-taking strategy. Journal of Educational Measurement, Vol 14, # 1, pp 15 - 22.
7.Frary, Bob (1988). Formula scoring of multiple-choice tests (correction for guessing). No. 3 in the series; Instructional Topics in Educational Measurement, B. S. Plake, Editor. Educational Measurement: Issues and Practices, 7(2), 33-38.
8.Johnson, N.L., Kotz, S., Balakrishnan, N. (1994) Continuous Univariate Distributions, Volume 1, Wiley. ISBN 0-471-58495-9 (Section 10.1)
9.Johnson, Norman L.; Kotz, Samuel; Balakrishnan, N. (1995). "Chapter 21: Beta Distributions." Continuous Univariate Distributions Vol. 2 (2nd ed.). Wiley. ISBN 978-0-471-58494-0.
10.Torres, C., Lopes, A., Babo, L. & Azevedo, J. (2011). Improving multiple-choice questions. US-China Education Review B 1, 1-11.
11.Taleb, Nassim Nicholas, (2001) Fooled by Randomness, Random House and Penguin.