MAKING THE GRADE:
A MATHEMATICAL MODEL OF GRADES AND GRADE GIVING

Dr. Alfred Carlson
Rose-Hulman Institute of Technology
Terre Haute, Indiana

1. INTRODUCTION

The evaluation of students (giving quantitative scores and grades) is a major aspect of engineering education, yet there is little information in the literature on mathematical theories of grading. In this paper I will discuss a mathematical model of the grading process that I am developing. The basic elements of the model are that grade getting is a stochastic process, involving the probability that a given student will be able to answer a question on a domain superimposed with differences in individual student's domain knowledge in the population, and variability (error) in the evaluation process itself. The model should be useful for developing more reliable and more uniformly accepted evaluation processes, for quantifying course improvement, and for understanding the quantitative aspects of grade scales.

For most educators, determining and assigning grades is one of the most difficult and least pleasant parts of their job. On the one hand, assigning a value to the knowledge a person has is a very inexact science, but on the other hand the grade has a lasting emotional impact on both the grade giver and grade getter. In addition, most experienced graders realize that there is imprecision in grade determination, and there is (at least for some) the nagging feeling that the student is getting the wrong grade, which may either under or over represent what they know about a subject or have learned during the learning period.

In this paper I am going to present a model that I think takes the first steps to understanding the grading process. The model necessarily is simplified, but has the advantage over the "touch and feel" approaches now used to decide what works in learning that it is quantitative and based on well established statistical methods. The model allows one to apply the same principles used in other areas of science to understand what is going on when we teach and how to quantitatively measure it.

A Definition of Knowledge

Part of the problem in grading is that most of us have a hard time deciding what knowledge (knowing something) really is. To get past this point, in this paper I am going to strictly define "knowing something" as being able to answer questions about that subject correctly. I am going to assume that grading is based on test scores (and not whether or not you like the person or think they are creative or ambitious, or lazy or irresponsible or any other reason you might come up with to give someone a certain grade). A measurement of a person's knowledge is strictly determined by testing that knowledge with questions that can be answered correctly or incorrectly.

When someone says you are knowledgeable they mean you know a lot about a certain area of information – the domain. We define domains when we offer a class about something or give a professional degree. Imagine the domain is "The State Capitals". If you were teaching this domain, what you would tell your students is that you expect them to learn all the capitals for all the states. In teaching them this domain, you might give them convenient ways to remember them, you might ask them to study a sheet, or you might teach them with a game. Regardless of how they learn the capitals, at some point you would give a test (or quiz) to determine how much they know on the domain you have defined.

There are basically two ways to approach testing domain knowledge. Knowing the names of all the state capitals is a pretty small domain and you could possibly simply give a test where all 50 are asked and a score is derived from that. Most domains are not so small and on a test you could not expect to ask the student to answer all of the possible questions on the domain. Instead you would ask a reasonable number of sample questions about the domain, more or less randomly selected. In the example of state capitals, you could, for example, select 10 states at random and ask the students to name the capitals of those 10 states. From the results you could infer something about the number of the state capitals the student actually knows. These two ways of testing, asking all possible questions or asking a sample set of questions will be examined in detail below.

2. A MATHEMATICAL MODEL OF TESTING

Testing is the process by which we try to estimate the inherent domain knowledge (IDK) of the person being tested. In particular, the IDK on the domain is the probability that the test taker will be able to answer a random question about the domain correctly. Suppose, for example, the domain is the "state capitals" and the test is: "What is the capital of [some state here]?" Suppose further that a person has been studying and knows the names of 74% (37/50) of the state capitals at the time he or she takes the test. The IDK is the probability (0.74) that the person will get the answer correct for a randomly selected state.

Figure 1 shows a schematic of the testing process. The point of the figure is that certain factors come between the IDK and the score someone gets when they are tested. At least five factors come between the IDK and test score. These are how the question is asked, the number of questions asked, the number of possible questions that can be asked, random error in question difficulty, and random errors in grading.

Factor #1 How the Question is Asked

How a question is asked influences the score a person gets. A question can be asked in an open answer way such as: "What is the capital of Arkansas?" The question could also be asked as a multiple choice question: "Which of these cities is the capital of Arkansas? a)

Lincoln, b) Matamoras, c) Little Rock. The question could also be asked in true false form as: “Matamoras is the capital of Arkansas” (True or False). For a person with a particular IDK, the chance they will get the answer correct depends on how they are asked.

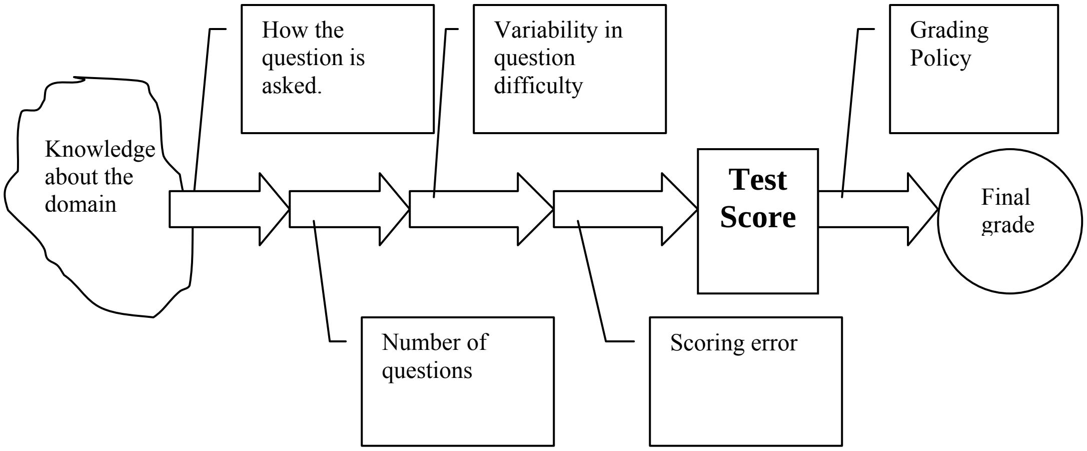
How the question is asked.
Variability in question difficulty
Grading Policy
Knowledge about the domain
Test Score
Final grade
Number of questions
Scoring error

Figure 1. Schematic showing the factors that separate how much a person knows about a domain from the test score and final grade they get. The individual factors are discussed in the text.

For an open answer question, the probability that a person will get the answer correct is simply equal to the IDK. For a multiple choice question, there are several pathways to get to the correct answer and these enhance the chances of selecting the right answer from the list given. The probability of getting the right answer on a multiple choice test is not equal to the IDK, only related to it. There will be a certain fraction of the answers that a person knows outright based on their IDK. There will also be a certain fraction of answers that a person would get right because they knew that all of the other answers were incorrect. The test taker would also get a certain number of questions right because they could eliminate all but two answers and then guess correctly between those. Finally the test taker could get some questions right by just guessing. The probability of getting a multiple choice question right I will call the apparent IDK (p_{app}).

Equation 1 expresses the value of the apparent IDK as a function of the actual IDK (p) in terms of the number of answers (n) given in the multiple choice form.

$$p_{app} = p + \Sigma \frac{n!}{(n-i)!i!} \frac{p^i (1-p)^{n-i}}{n-i} \qquad (1)$$

i is an integer less than or equal to n. The relationship between p_{app} and p is shown in Figure 2 for different numbers of potential answers given. A person with an IDK of 0.5 should get 80% of the answers correct on a 3 answer multiple choice test etc.

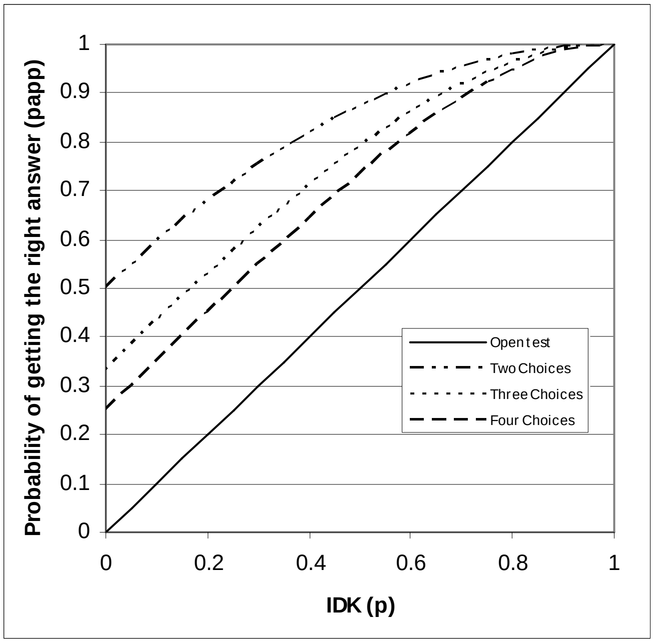

Figure 2. The apparent IDK is higher than the true IDK when multiple choice questions are posed to the test taker.

Factor #2 – Number of questions asked

The number of independent questions used to determine an IDK is also important. Suppose N is the number of independent domain questions that could be asked on a test and you are asking n on the particular test (n ≤ N). If N is very large compared to n (strictly infinite) then the probability that the student will get r questions right out of the n asked (score = r/n) will be given by the binomial distribution (Equation 2):

$$P_r^n = \frac{n!}{r!(n-r)!} p_{app}{}^r (1 - p_{app})^{n-r} \tag{2}$$

p_{app} is the apparent IDK of the student, the probability that he or she will get one particular question right on a particular type of test. The most probable score (r/n) will be equal to p_{app}. However, there is a high probability that the student will get less than or more than the most probable score because of the chance nature of the test. In fact, if $p_{app} = 0.7$ on a 10 question test, 70% of the time the student will either get 6 or less right or 8 or more right. Paradoxically, you should not expect a person with an apparent IDK of 0.7 to get 7/10 right on such a test!

The standard deviation is a measure of the variation in the scores a person would be expected to get for a particular IDK. For the binomial distribution, the standard deviation (σ_{nq}) due to the fact that a finite number of questions is used to determine the IDK is given by (Equation 3):

$$\sigma_{nq} = \sqrt{\frac{p_{app}(1-p_{app})}{n}} \tag{3}$$

More questions lead to smaller error in determining the IDK from the test score. This effect will be looked at with a computer experiment in the Results section.

Factor #3 Number of Possible Questions

Sometimes the number of possible questions (N) is not much greater than the number asked on a test. In this case the hypergeometric distribution (Equation 4) represents the probability of a student getting r questions right on an n question test (score = r/n).

$$P_r^{n,N} = \frac{\left[\dfrac{p_{app}N!}{r!(p_{app}N-r)!}\right]\left[\dfrac{N(1-p_{app})!}{(n-r)!(N(1-p_{app})-r)!}\right]}{\left[\dfrac{N!}{n!(N-n)!}\right]} \tag{4}$$

The most probable value for r is one integer value below np_{app}. The difference between the hypergeometric distribution of scores and the binomial distribution is that the hypergeometric distribution is narrower than the binomial distribution. It is more likely for a student to get a score more reflective of her apparent IDK in tests in which the number of questions asked is closer to the maximum number that could be asked, hence one will get a more precise measurement of IDK from a domain that is smaller in terms of the number of possible questions than from one that is not. If one asks all possible questions (n = N) the test taker will get a score that is exactly equal to the value of p_{app}.

The standard deviation because of the number of questions asked for the hypergeometric distribution (Equation 5) is given by:

$$\sigma_{nq}^{hg} = \sqrt{\frac{p_{app}(1-p_{app})}{n}\frac{(N-n)}{(N-1)}} \tag{5}$$

Figure 3 compares the probability distribution for the binomial distribution and hypergeometric distribution when 25 questions are asked. In the case of the hypergeometric distribution the total number of possible questions was assumed to be 50.

Factor #4 Variability in the difficulty of the questions asked

It is extremely difficult to make up a test in which all of the questions of equal point value are really equally probable to answer correctly. When the value of p_{app} varies from question to question, this adds further random error to the test score. The "question difficulty" error is additive to the error (standard deviation) caused by the number of questions so that (Equation 6),

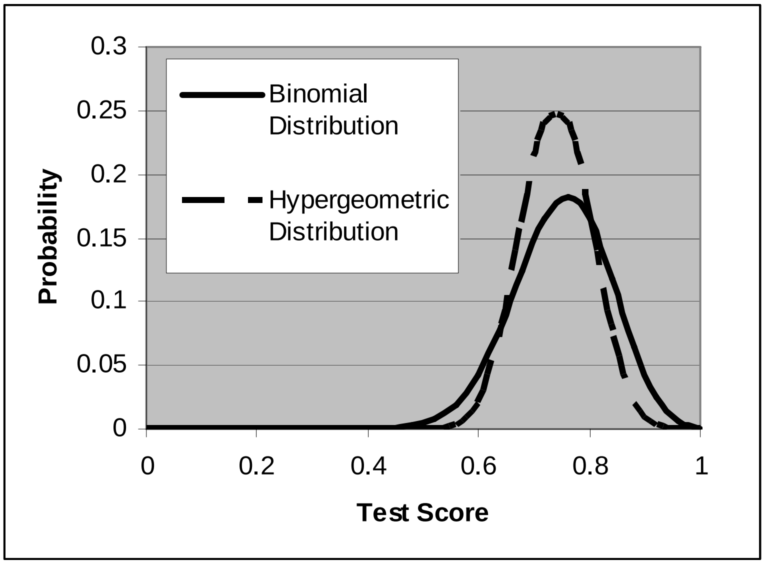

Figure 3. Comparison of score probability distribution predicted by the binomial (solid line) and hypergeometric (dashed line) distributions. n = 25 questions were asked and the IDK of the student is assumed to be 0.75. The number of possible questions (N) was assumed to be 50.

$$\sigma_{test} = \sqrt{\sigma_{nq}^2 + \sigma_{qd}^2} \qquad (6)$$

is the total "test error" caused by the number of questions and inability to make them all the same difficulty. σ_{nq} represents the error associated with the number of questions and σ_{qd} represents the error associated with the question difficulty.

Factor #5 Grading Error

The fifth factor is grading error. For tests in which the grader must evaluate the work, random errors creep into the score that are caused by mistakes or inconsistencies in grading (σ_{gr}). Again this kind of error is additive to the error caused by the number and difficulty issues of the questions. The overall "testing" error includes grading error,

$$\sigma_{testing} = \sqrt{\sigma_{nq}^2 + \sigma_{qd}^2 + \sigma_{gr}^2} \qquad (7)$$

In the Examples section below, computer experiments are used to show how these errors influence an individual student's score on a test.

3. MODELING RESULTS – ILLUSTRATIVE EXAMPLES

Example 1. Influence of how the question is asked

Suppose that a medical student is learning the use of various therapeutic drugs. The assignment is for the student to learn the purpose of 500 drugs. Suppose also that a particular student has learned 75 % of the drugs (375) at the time of the test. The student's domain knowledge is determined with a test.

One way to set up the test (as an open answer test) is to simply list all 500 drugs and leave a space for the student to write down the purpose of the drug. A student with a 0.75 IDK for this domain (75%) would receive a 75% on this kind of test.

Another way to set up the test is to make the test "multiple choice" in which several possible answers are given after each drug. A student with an IDK of 0.75 should get better than a 75% on a test like this. Barring ambiguities, the student would get 375 answers correct because she would immediately know this many answers. The student should be able get another 70 answers correct by knowing that 2 of the answers to a given question are wrong, thereby indicating the third answer must be correct. The student should be able to get an additional 23 questions correct, by knowing that one answer is incorrect and guessing between the other two. Finally, the student should be able to essentially guess correctly on about 3 questions, even though she cannot eliminate any of the answers. The student should be expected to get (375 + 70 + 23 + 3 = 471 ---94%) on a 3 question multiple choice test even though her IDK is only 0.75.

Example 2. Influence of number of domain questions and questions asked

Using the equations presented above, a "computer model" of an "individual student" was built in which results were randomly generated consistent with statistically valid behavior using an assigned inherent domain knowledge (IDK) value. In other words, stochastic predictions of an individual student's anticipated scores were made with the program(s). The computer model therefore allows one to predict realistic "life-like" behavior (results) when a student experiences the random influences of which questions might be asked about a domain, variations in the difficulty of the questions asked, and grading errors. The advantage of the model is that it can show one how small numbers of questions can generate erratic testing results and how changes in the number of questions, size of the domain, question error, and grading error can influence the ability to predict the IDK with a test. The computer model, as opposed to "statistical models" can predict results in the "short run" rather than in the "long run".

In terms of computations, the model works as follows. First, the user assigns an IDK to a student and the computer generates expected single test results for that student (based on that student's IDK) when n questions are asked out a domain with N possible questions. The program first calculates the standard deviation based on Equation 5. Using this standard deviation and a randomly generated probability value between 0 and 1 (from the

Excel function RAND()) the program calculates a statistically consistent score from the NORMINV function

Suppose the test giver decides to try to estimate the student's drug IDK by giving one of two open answer tests asking less than all 500 drugs. The test giver is trying to determine the quantitative differences between the two options of either 20 or 200 drugs being selected at random and asked on the test.

If 20 questions are asked, a test taker with an IDK of 0.75 would "most probably" get 15 of the drugs correct, but there would also be a significant probability that the student would get some other number of questions correct. The statistical variation in individual test results when 20 questions are asked out of a domain of 500 is shown in Figure 4. A statistically valid set of repeat scores for 10 retakes of the same test is shown in Figure 5.

Also illustrated are the statistical distribution (Figure 4) and variation expected (Figure 5) when the number of drugs asked is 20 or 200. Increasing the number of questions on the test increases the precision of the IDK measurement.

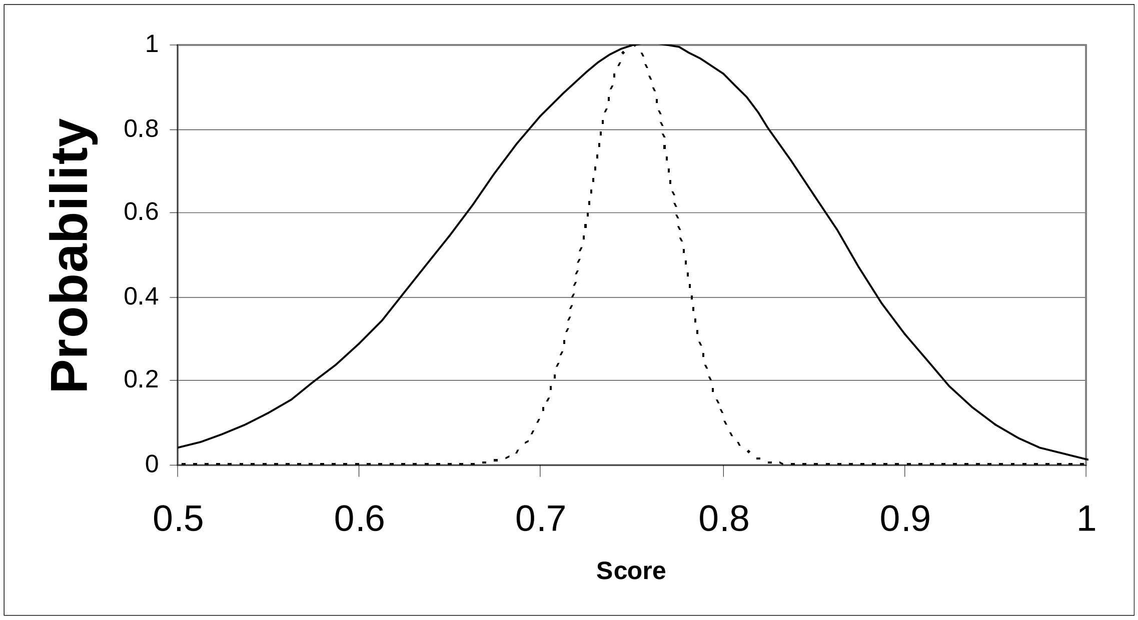

Figure 4. Comparison of score probability distribution predicted when 20 (solid line) or 200 (dotted line) questions are asked and the IDK of the student is assumed to be 0.75. The number of possible questions (N) was assumed to be 500.

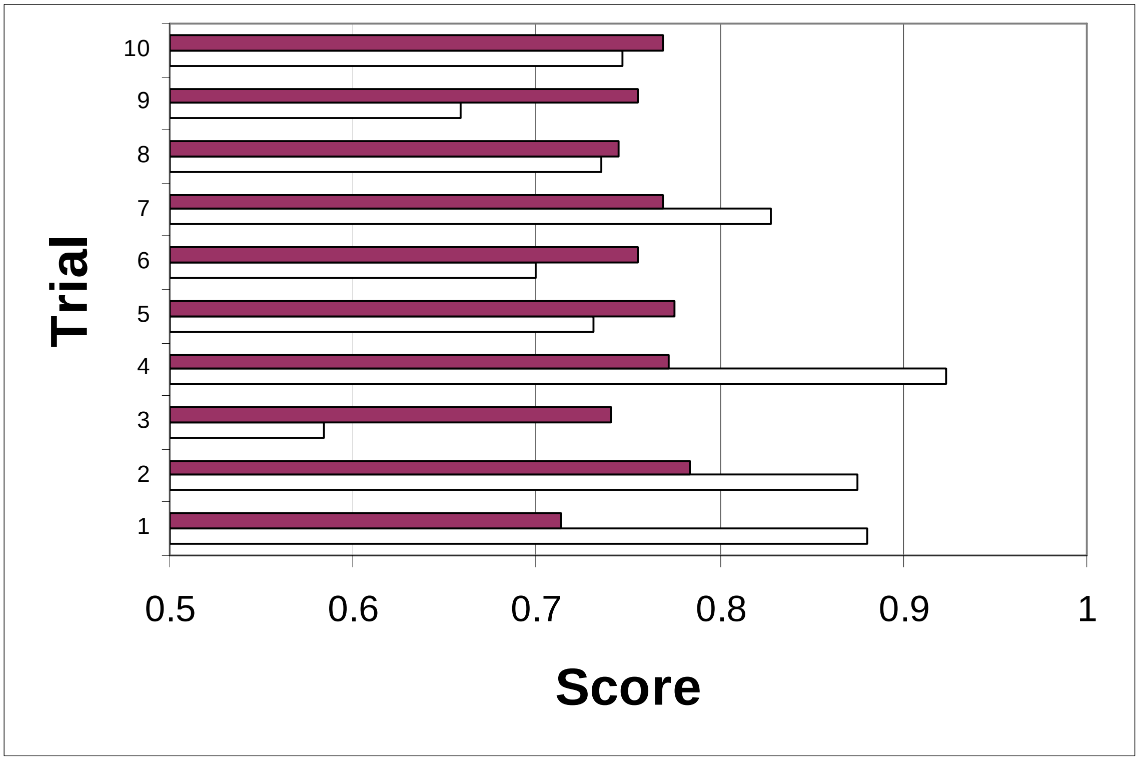

Figure 5. Variation of scores expected when 20 (open bars) or 200 (solid bars) questions are asked and the IDK of the student is assumed to be 0.75. The number of possible questions (N) was assumed to be 500.

Figures 4 and 5 illustrate the difficulty in determining the true IDK when a small number of questions are asked. On a 90, 80, 70, etc. scale, A student with a solid "C" IDK could easily fail a single test (score < 0.60 – trial 3) or get an A (score > 0.90) on any given day if only 20 drugs (questions) were asked. On the other hand, asking 200 questions basically ensures that the student will get between 70 and 80, in this case a C.

Example 3 . Influence of variation in question difficulty and grading error

Even if a large number of questions relative to the size of the domain are asked, uncompensated variation in the difficulty of individual questions and grading errors can cause the estimation of the student IDK to be imprecise. Suppose for example that some of the drug actions the students are expected to memorize are more difficult than others for some reason. This introduces variability into the value of the probability that a student will get that definition correct. In this case suppose from drug to drug there is a 5% standard deviation in the probability that a student will get the answer to a given drug function correctly memorized. Within one standard deviation, the"IDK" of the student for a given drug would vary from 0.7 to 0.8. In addition, suppose that the ability to correctly score an answer to a given drug function had a standard deviation of another 5%.

Figure 6 shows a typical set of statistically valid set of test scores when 200 questions are asked but when there is question difficulty variation and/or scoring error. In this case there is a significant probability that the test score will not be representative of the IDK (0.75). In terms of letter grades for 90, 80, 70 scoring, the student would have gotten an inflated grade (B or better) 3 of 10 times simply due to random chance. Other examples would show the student could frequently get an undeserved lower score.

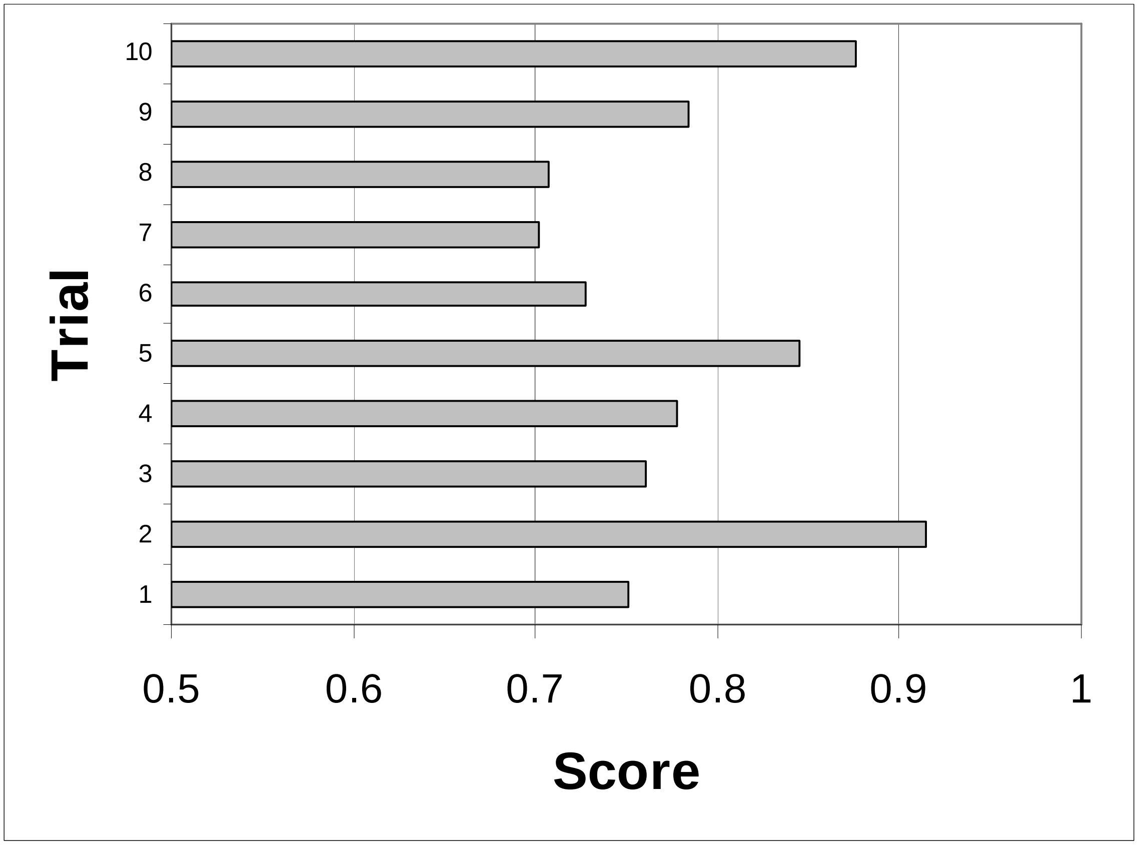

Figure 6. Variation of scores expected when 200 questions are asked and there is a 0.05 standard deviation in question difficulty and a 0.05 error in grading the test. The IDK of the student is assumed to be 0.75. The number of possible questions (N) was assumed to be 500.

4. CONCLUSIONS

The analysis presented here shows that multiple choice questions inflate the apparent knowledge of the test taker. In addition, it demonstrates that testing error, which includes the number of questions that are asked on a domain (relative to the number of possible questions), variations in the difficulty of the questions, and scoring error, can result in test scores that are not consistently representative of the knowledge of a subject a student really possesses.

References: Various statistical concepts found in standard textbooks, in the Excel analysis functions, or on the internet were used in this study.