article:1266

Student Assessment Precision in Mechanical Engineering Courses

When experimentalists perform measurements, there are rigorous ways to establish the accuracy of those measurements. Scientific publications include information on the accuracy of the measurement. In contrast to scientists, who quantify the uncertainty in their measurements, educators almost never quantify the uncertainty in their measurement. Students do not as a general rule receive report cards with grades such as Math 101: 80%±5%. This study was undertaken to attempt to quantify the uncertainty in the grading of students. Owing to the absence of a "gold standard" for the measurement of student ability, it is not possible to determine the accuracy in one's measurement of student ability. However, one may determine the precision of the measurement by repeating the measurement numerous times and observing the variability in the repeat trials. This variability was determined by studying student performance on mid term tests and/or quizzes and final examinations in six core undergraduate courses offered by the Department of Mechanical Engineering at UBC. If the precision associated with each of the mid term and final exam marks was high, the correlation between the two marks should also be very high because it is common for 50% or more of the material tested on a final exam to be the same as that tested on a mid term, and much of the remaining exam material may build on concepts tested in the mid term.

Statistical comparison of the grades of students on these two examinations shows only a modest correlation between grades, with correlation coefficients between r=0.4 and r=0.7. A scatter plot of individual grades on mid terms and final examinations resembles a shotgun target for which the target has been inclined. The trend of the points is upwards, but the scatter is very large. In practice, the root mean square deviation between the marks a student receives on quizzes and the final examination is between 10% and 15% for most courses, even if the quizzes and final examination have comparable means and standard deviations.

The modest correlation between quiz and final examination grades may be attributed to several factors: inaccurate grading of examinations, differences in the material tested on the mid terms and finals, students having an “off day,” and “sample size error.” We may rule out the first explanation because repeat grading of examinations by different instructors yields only modest differences in assigned grades (when the grades assigned by different instructors are appropriately normalized). We may likewise eliminate the second possible explanation. For both mid terms and final examinations, the successful solution to a question typically involves a mixture of physical insight and mathematical manipulation. The questions yield a well-defined correct answer within half to one page of work. In terms of Bloom’s taxonomy, the skills required for success in such examinations are almost exclusively cognitive. Specifically, the application of knowledge forms the basis of most examination questions, with some weight given to student comprehension of course material. The similar character of mid term and final examination questions in these courses makes it doubtful that this is the source of the grade variability. One might hypothesize that the variability is due to students having an “off day.” This hypothesis was refuted by showing that the variability in student performance between different parts of an exam held on a single day is even larger than that between mid terms and final

examinations.

The final possible cause of the modest correlations requires additional explanation. A typical engineering course might involve 10 key concepts. An average student might understand 6 of the 10 concepts. Owing to time constraints, a typical exam might test a student's understanding of only 5 of the 10 concepts. An average student, then, could theoretically get anywhere from only 1 of the 5, to all 5 of the 5, questions correct, based purely on the chance decision made by the instructor to ask a particular set of questions compared to an equally valid different set of questions. Although it is difficult to prove conclusively, qualitative arguments such as these imply that this sampling error is the primary source of the low correlations.

Whatever the sources of the modest correlation, the implication is the same -- the precision of student assessment in the courses studied is only approximately ±5-10%. The implications of this statement for "yes-no" decisions such as failure in a course and the awarding of scholarships is significant.

Author 1:Sheldon Green email: green@mech.ubc.ca

: Back to Fall 2005 Issue Vol. 1, No. 2

: Back to List of Issues

: Back to Table of Contents