The Precision of Grades in Physics Courses

David M. Harrison[a]
Department of Physics
University of Toronto

Almost all teachers are required to submit final marks for their students. In this note we discuss the uncertainty in those marks with an emphasis on marks in physics courses. We will estimate that the statistical uncertainty in final marks is at least 4% out of 100%, and is probably significantly higher than this. The implications of this result on things like calculating Grade Point Averages are large.

For multiple-choice tests there is a large body of statistical work[1], which we will briefly review. The reliability r of a particular test can given by the Cronbach α coefficient:[2]

$$r = \alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i(1-p_i)}{\sigma^2}\right] \tag{1}$$

where K is the number of questions on the test, p_i is the fraction of correct answers for question i, and σ is the standard deviation of the scores on the test. The values of r are between 0 and 1. Professionally developed high-stakes standardised tests achieve reliabilities of at least 0.9, and by convention values of r less than 0.5 indicate a poorly designed test. It turns out that the reliability of a test increases as the number of questions on it increases.[3,4,5]

From the reliability, the standard error or measurement s can be calculated.[6] This is the statistical uncertainty in each individual student's mark on the test, and is given by:

$$s = \sigma\sqrt{1-r} \tag{2}$$

The interpretation of s is similar to the standard deviation of experimental measurements: 1 s corresponds to a 68% confidence interval, 2 s to a 95% confidence interval, etc.

For physics tests the issue of using multiple-choice questions as opposed to long-answer questions, which are marked in detail with part marks awarded, is religious, and we will try to avoid those arguments here. In our large (900-student) 1st year university physics course primarily for life-science students, we typically have about 75% of the marks on each test and exam determined by multiple-choice questions, and about 25% determined by one or more long-answer questions. Since our typical multiple-choice question takes the student about 5 minutes to do, we can have between eight and sixteen such questions on each test and exam. This is different from tests in subjects that are fact-based, such as

[a] Email: david.harrison@utoronto.ca

introductory psychology, where colleagues in those Departments report that each question on tests and exams in their introductory course takes the students about 1 minute to do, so in the same time period up to 50 questions can be asked.

Over the past couple of years the best reliability we ever achieved on the multiple-choice section of our tests and final exam was $r = 0.70$ with a corresponding standard error of measurement $s_{\mathrm{MC}} = 11\%$. Thus we can only distinguish between marks on this part of the test to within an uncertainty of over a full letter grade. The comparatively poor reliability and corresponding high error of measurement is undoubtedly in part because of our lack of skill in setting a good test, but it also a reflection of the small number of questions we can ask. Courses in, say, introductory psychology, with more multiple-choice questions, often achieve higher reliabilities and smaller errors of measurement than we have managed to achieve. Below we will assume this best value of $s_{\mathrm{MC}} = 11\%$ for all tests and exams in a model course. Therefore, our calculation of the uncertainty in the final grade is definitely a lower bound.

For the long-answer section of our tests and exam, we do not know of any way to estimate a standard error of measurement s_{LA}. For our typical test with 75% multiple-choice and 25% long-answer marks, the test mark is:

$$\begin{aligned}\text{test mark} = (\text{multiple-choice} \pm s_{\mathrm{MC}}) \times 0.75 + \\ (\text{long-answer} \pm s_{\mathrm{LA}}) \times 0.25\end{aligned} \tag{3}$$

Since the values of s are errors of precision, they should be combined in quadrature, i.e. the square root of the sum of the squares. Thus the uncertainty in the test mark s_{T} is:

$$s_{\mathrm{T}} = \sqrt{(s_{\mathrm{MC}} \times 0.75)^2 + (s_{\mathrm{LA}} \times 0.25)^2} \tag{4}$$

If we assume that the multiple-choice and long-answer sections have equal standard errors of measurement, $s_{\mathrm{MC}} = s_{\mathrm{LA}} = 11\%$, i.e. that they both are equally effective in assessing students, then from Eqn. 4 $s_{\mathrm{T}} = 8.70 \cong 9\%$. We will also assume that the standard error of measurement on the final exam, s_{Final}, also is 9%.

The uncertainty in the test mark given by Eqn. 4 is not highly dependent on the uncertainty in the long-answer section. Figure 1 shows the value of s_{T} for values of s_{LA} from 1% to 20%. The value of s_{T} varies from 8.25 to 9.65%. In the figure, the horizontal line is the value of Eqn. 4 for the assumed values of $s_{\mathrm{MC}} = s_{\mathrm{LA}} = 11\%$.

We will also assume that for the parts of the final mark in the course that are not from tests and the exam, such as problem sets or laboratories, there is no error of measurement in these marks, i.e. that they are perfect in assessing students; this assumption is certainly not true.

Often in our course, we have two term tests and a final exam. Each test counts for 15% of the mark in the course and the final exam counts for 40%. Therefore, using our extremely optimistic assumptions the final mark in the course is:

$$\begin{aligned}\text{final mark} &= (\text{Test1} \pm 9) \times 0.15 + (\text{Test2} \pm 9) \times 0.15 + \\ &\quad (\text{Final} \pm 9) \times 0.4 + (\text{term work} \pm 0) \times 0.30\end{aligned} \tag{5}$$

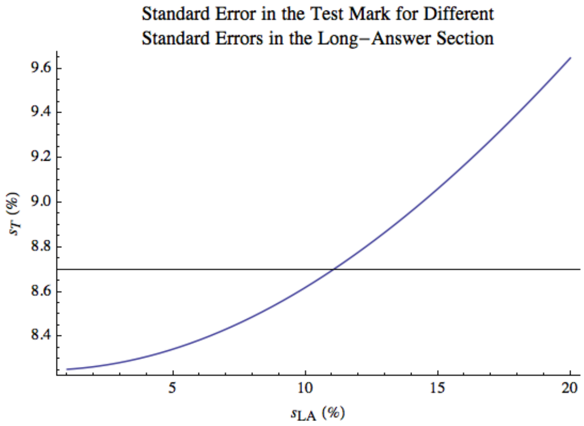

Figure 1. Dependence of the standard error of the test mark on the standard error of the long-answer section

The standard error of measurement for the final mark, s_{FM}, is:

$$s_{\mathrm{FM}} = \sqrt{(s_{\mathrm{Test1}} \times 0.15)^2 + (s_{\mathrm{Test2}} \times 0.15)^2 + (s_{\mathrm{Final}} \times 0.4)^2 + (0 \times 0.3)^2} \tag{6}$$

Since we have made many assumptions that make these errors lower bounds, the actual uncertainty in the final mark is certainly larger than the value obtained by just propagating the errors in Eqn. 6. Therefore the result of the calculation gives us:

$$\Delta(\text{final mark}) > 4\% \tag{7}$$

Thus final marks of, say, 76% and 77% are the essentially identical within errors.

The dominant contribution to the result Eqn. 7 is from the uncertainty in the final exam. If we model a course with one term test counting for 30% of the mark in the course, with the final exam still counting for 40%, the uncertainty in the mark rises to $\Delta(\text{final mark}) > 4.5\%$.

We made many assumptions to get to Eqn. 7. So the calculation is a type of Fermi Question: different sets of reasonable assumptions will lead to a very similar result. Therefore we believe that the uncertainties in final marks in our courses are probably comparable to those given to students in most physics courses at most schools.

At the University of Toronto, a 76 corresponds to a letter grade of B and a 77 corresponds to a letter grade of B-plus. For calculating a Grade Point Average (GPA) the university makes a distinction between B and B-plus with the former having a value of 3.0 and the latter 3.3. So the effect on the student's GPA of these two essentially identical final marks is large. This same ill-advised procedure is common in one form or another at many schools.

In Toronto we have had considerable discussion about what to do about this, but without a satisfactory resolution. For example, we could convert grades of 76 to 77. But then what about 75? And if we change 75 to 77, then what about 74? We have also considered rounding all marks to the nearest 5, but that would mean that a mark of 78 goes to 80, an A, while 77 goes to 75, a B. We also discussed rounding up to the nearest mark that is evenly divisible by 5, but this makes a huge distinction between a 75, which stays the same, and a 76, which goes to an 80. Perhaps the only resolution is to drop the GPA calculation entirely.

Failing that institutional change, when we are confronted with the list of student names and final course marks that we are to turn in at the end of the course, we need to at least be sensitive to the large uncertainty in the numbers.

Last revision of this document: December 9, 2013

[1] See, for example, L.M. Harvill, "An NCME Instructional Module on Standard Error of Measurement," ITEMS: Instructional Topics in Educational Measurement (1991), http://ncme.org/linkservid/6606715E-1320-5CAE-6E9DDC581EE47F88/showMeta/0/ (retrieved June 4, 2013), or AERA, APA and NCMD, *Standards for educational and psychological testing* (American Psychological Association, Washington D.C., 1985).

[2] L. Cronbach, "Coefficient alpha and the internal structure of tests," Psychomerika **33** (1951), 297. There are other statistics that measure the reliability, such as the Kuder-Richardson Formula 20, but for our purposes they are all equivalent.

[3] J. Nunnaly and L. Bernstein, *Psychometric Theory* (McGraw-Hill, Toronto, 1994).

[4] R. Cohen and M. Swerdlik, *Psychological Testing and Assessment* (McGraw-Hill, Toronto, 2010).

[5] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," International Journal of Medical Education **2** (2011), 53.

[6] W.G. Mollenkopf, "Variation of the standard error of measurement," Psychometrika **14**, (1949), 189.