

An actual level of difficulty of test problems and its subjective perception by students

Dr. Valentin Voroshilov, Physics Department/ Boston University, valbu@bu.edu
2009 AAPT National Winter Meeting / Chicago, IL

Developing tests (midterms and finals) is an important part of a job every instructor has to do. The transition from Introductory to Upper-Level Courses is difficult sometimes because of a significant increase in the difficulty of tests.

Should a test be easy, difficult, how do we measure difficulty of the test, how do students decide if a test was difficult, should a faculty even ask these questions?

Assessment process is very subjective, but it is a fundamental part of teaching, since, in a way, it shows how good we are at it. To bridge our understanding of *what* should be assessed and *how* it should be done, we conducted a survey, asking students to evaluate how difficult the test problems was from their point of view and comparing the students responds with (a) average grade for each problem (shows its actual difficulty), and (b) instructor's opinion.

Preliminary analysis shows that students' evaluation of the difficulty of a test problem, as well as provided by a faculty, often does not correlate with the actual difficulty level.

At this point the first stage of the research still is in development. We are trying to understand how students see problems offered to them on a test.

The scale offered to students had three possible options, i.e. 1 (the problem was easy); 2 (the problem was not too easy but not too hard); 3 (the problem was hard).

For students, ranking test problems was not a requirement and did not have any reflection on the grade for the test.

Still, we were pleased that many students have responded, and not all of the responds were pointing at the highest level of difficulty. It tells us that many students appreciate our attempt to address difficulties students may have experienced when taking the course.

At the end of a test a spreadsheet is prepared which includes test results as well as ranking of the problems of a test. We calculate an average grade for each problem, an average rank for each problem and also an adjusted average rank, which excludes from consideration students who did not rank a problem. Since the most of the students ranked the problems the adjusted average rank is practically equal to an average rank.

Then we calculate a relative actual difficulty and relative perceptive difficulty for each problem.

A relative actual difficulty of a problem is calculated as a ratio of the average value for the problem to the maximum grade assign to the problem. A relative perceptive difficulty of a problem is calculated as a ratio of the average value of the problem's rank to the maximum rank (which is equal to three in our case).

Low value for the relative actual difficulty of a problem means a relatively small number of students had the problem completely solved. Low value for the relative perceptive difficulty of a problems shows that most students saw the problem as relatively easy. Problem #1 has its relative actual difficulty of 0.61 (it was hard), and the relative perceptive difficulty is 0.84 (students saw it as a hard problem).

The difference between a relative actual and relative perceptive difficulties is: (1) a relative actual difficulty lies between 0 and 1; a relative perceptive difficulty lies between 1/3 and 1; (2) for a relative actual difficulty 0 represents the hardest problem; for a relative perceptive difficulty 1/3 represents the easiest problem.

To match the two measures we can calculate a true relative perceptive difficulty of a problem. The true relative perceptive difficulty (Y) can be found through the relative actual difficulty (X) as $Y = [2*(1 - X) + 1]/3$. This simple linear function just matches in a simplest way the easiest actual difficulty (1) with the easiest perceptive one (1/3), and the hardest actual difficulty (0) with the hardest perceptive one (1).

Now we can calculate the true relative perceptive difficulties for the problems and compare them with assigned by students. The result is presented in the table 1.

| | P1 | P1/r | P2 | P2/r | P3 | P3/r | P4 | P4/r | P5 | P5/r | P6 | P6/r | P7 | P7/r | P8 | P8/r |
|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|
| Rel | 0.61 | 0.84 | 0.90 | 0.54 | 0.78 | 0.83 | 0.80 | 0.74 | 0.85 | 0.63 | 0.76 | 0.79 | 0.69 | 0.71 | 0.87 | 0.67 |
| True | | 0.92 | | 0.49 | | 0.66 | | 0.63 | | 0.56 | | 0.69 | | 0.79 | | 0.52 |
| | | < | | | | | | | | | | | | < | | |

The bold font represents true relative perceptive difficulty. When a true relative perceptive difficulty is grater than a relative perceptive difficulty assigned by students that means students on average underestimated the problem, the problem turned out to be more difficult than students thought. Table 1 has only two problems like that, namely, problem 1 and problem 7 (“<” shows those problems).

When a true relative perceptive difficulty is less than the one assigned by students, that means on average students overestimated the problem, i.e. they saw it as more difficult than it was. We see that most of the problems (6 out of 8) were the ones where students overestimated the difficulty.

We will call “scenario 1” the situation when students overestimate more test problems than underestimate ones.

More statistics is needed to see if this situation happens most of the time, or the opposite case (“scenario 2”, when students underestimate more problems than overestimate) is also of a common practice.

From our point of view it is better for the psychological climate of a class if scenario 1 is realized, since students feel more confident after a test and believe that their hard work over the course pays back.

Another statistics, which is a point of our interest, is how faculty assesses the difficulty of problems used to test students. This work is at the very beginning stage (preliminary data show that faculty leans to underestimate the difficulty of problems).

The author wants to express his gratitude for fruitful discussions to Prof. Andrew Duffy (BU), and Prof. Ophelia Tsui (BU).