Use of Web-based Testing of Students as a Method for Evaluating Courses

Harry Roy
Department of Biology
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: royh@rpi.edu
FAX: (920) 465-2769

Abstract: Educational innovation, like other educational processes, requires a tool for measurement of its effect on student learning. This is especially important when it involves new techniques. The testing of students both before and after a class (pre- and post-testing) is an established method for measuring student progress. This is useful because it normalizes the results by effacing some of the variability in preparation of the students. Automated test administration and grading makes it much easier to implement pre- and post-testing. I used the WebCT quiz tool to administer pre- and post-tests to students in 5 successive semesters of a course in Genetics and Evolution, and I collected data on course grades for 12 semesters of the same class, containing from 10 to 60 students each. Students showed a gain in learning $g$ ranging from $0.54 \pm 0.32$ to $0.88 \pm 0.21$. Course grades were weakly negatively correlated with class size ($r=-0.37$). Based on these results and published literature, I suggest that the gain in learning will vary with discipline, teaching technique, and student background, and that automated pre- and post-testing could be useful for comparisons between different instances of the same class containing large numbers of students.

Keywords: computerized testing, gain in learning, automatic assessment, evaluation

Introduction
Evaluating courses is important. One can evaluate the performance of students by giving tests and one can evaluate the reactions of the students to courses by soliciting their evaluations. But class size, limits on faculty time, the variability among students and faculty, and other factors make educational experiments hard to design, hard to implement, and hard to interpret when completed. In a dynamic institution, there will be new ideas proposed for teaching. But skeptics both on the faculty and in the administration may ask, "Is it worth the effort? Do students really learn better with these methods?" The answer will affect the allocation of resources for improving instruction at our university and at many others.

Robert Hake (1998) introduced a measurement called $g$, the gain in learning, which was calculated from the results of pre- and post-tests of about 6000 college physics students at many universities in the United States. This study documented that American students show about a 20% improvement in performance on tests of basic physics concepts upon completion of a year of college physics, taught in the traditional manner, as characterized by those faculty supplying the data. (A startlingly low figure!) Cummings et al. (1999) used this measurement to evaluate the effectiveness of studio physics, as compared with interactive lecture demonstrations. (Studio physics is an interactive learning approach that integrates lecture, recitation, and lab in a single classroom; interactive lecture demonstration is a method that requires students to make predictions about the results of demonstrations, view the demonstrations, and then draw lessons from the comparisons). The results of testing students on basic physics concepts yielded a value of $g = 0.35$ for interactive lecture demonstrations. This was significantly higher than that obtained either for studio physics or for standard classes, consisting of lectures,
labs, and recitations ($g=0.2$). It would appear from this that interactive lecture demonstrations provide the most effective learning situation; but what interests me here is the effort involved in getting this kind of information.

I use the WebCT quiz tool to provide automatically graded homework assignments to my students in Genetics and Evolution, and to administer term examinations. Since I teach the class using a studio format, with a limit of 25 students per section, I have to teach it several times a year in order to handle all the students enrolled each year. Learning of the work of Cummings et al. (1999), it occurred to me that I could easily obtain data on the effectiveness of my teaching basic concepts in Genetics and Evolution, from one semester to the next. I was curious about the magnitude of the gain in learning $g$ that I might expect. Also I wondered if this technique might provide an objective measure that could be compared from one semester to the next, as opposed to student evaluations. Rightly or wrongly, student evaluations seem to so many of us an unreliable way to evaluate teaching. The results indicate to me that quantitative evaluation is marginally useful for small classes, but could be very useful for large ones.

**Methods**

I had already set up a database of genetics and evolution questions, including multiple choice, matching, and problem solving questions, using the WebCT software (http://www.webct.com). The multiple choice questions included many with different values assigned for different answers, depending on the errors that were likely to produce them. The matching questions allow a longer list of answers on one side, so that simple elimination is not sufficient for completing the question. The calculation questions generate slightly different versions of the same question for each student, and allow the instructor to specify limits for an answer, which is a big advantage over more traditional fill-in questions that require an exact match. From this database, I chose representative questions on each of 15 topics, questions that covered basic concepts in the field. Each member of a pair of questions was judged to be of equivalent difficulty. I assigned a member of each pair to either the pre-test or the post-test. I weighted all questions equally, and each test had a total of 100 points. I modified the tests from one offering of the course to the next by replacing questions with others of comparable difficulty. After training the students in the use of the software, I administered the pre-tests on the first day of class. I told the students that they were expected to do their best, but that the examination was designed to set a baseline for the post-test, not to have an effect on their grade. I allowed plenty of time for each exam; and few students used all the available time. I calculated the value of $g$ for each student by the formula: 

$$
\frac{\text{[(Post-test)-(Pretest)]/100}}{\text{[(Pre-test)]}}
$$

In other words, the measurement $g$ assesses the extent to which each student improves. This testing regime was implemented for five
successive presentations of my sophomore Genetics and Evolution course.

**Results and Discussion**

Using pre-tests and post-tests that were graded automatically using WebCT, I obtained the results shown in Table I for my class in Genetics and Evolution. The parameter \( <g> \), the gain in learning, ranged from 0.54 ± 0.32 in the first attempt, to a high of 0.88 ± 0.21. In experiments with interactive lecture demonstrations in physics, a lower gain in learning \( (<g>=.35) \) was reported (Cummings et al. 1999), but it was significantly higher than that reported for traditional physics classes that did not use interactive learning techniques \( (<g>=.2) \) (Hake, 1998). The absolute values of the gain in learning reported for my course are probably not directly comparable to those reported for physics (Cummings et al. 1999). One possible reason is that in their study, the post-test did not count toward the students' grades, and thus, they may not have prepared as well as mine did. Another possible reason is that the two subjects may differ in the ease with which the subject matter may be learned. The usefulness of this testing technique may lie in its ability to detect changes in the effectiveness of the same basic course when taught at different times, by different instructors, or using different pedagogical methods. For example, it seems as if the course taught in the spring produced a lower learning gain despite the small class size. A T-test suggests that at a level of \( p=.05 \), the spring class is not just a subset of the fall class (not shown). What could be the reason for this? Well, the spring edition of the class has a higher proportion of non-majors (not shown here). The ease with which such students may learn the material could be less than that of biology majors. On the other hand, teaching assistants were provided for the fall course, but not for the spring course. This could be a crucial factor. The point is not to explain the difference here, but to show that data obtained this way can be used to formulate hypotheses.

The use of automatic grading of standardized questions avoids objections that the grading would be biased by differences among the instructors. Also, having a measurement that is so easily obtained, and that relies on objective criteria, could be a useful supplement to the commonly used student evaluations, which are widely regarded with suspicion by faculty. One important factor is the amount of data on students, both for objective tests and for student evaluations. The sensitivity of any method will depend on the numbers of students enrolled. Small numbers of students yield high standard deviations, enabling the detection of little but the greatest differences. The evaluation of small classes thus remains problematical, regardless of the methods used.

A large number of values of \( <g> \) for different size classes was not available, but I did have class averages for 12 offerings of the same course taught to different numbers of students over a 3-year period. These classes covered the same material, used similar test questions, and had, for the most part, similar populations of students. The class averages were based on homework, examinations, and term papers. The data in Figure 1 show a slight negative correlation \( (r=-0.37) \) with class size. When two small classes of highly select pre-medical students were pulled from the data set, there was still a correlation of \( r=-0.32 \). Student performance increases, as measured by grades, with smaller class size. This is no surprise. However, the scatter of the data is great, indicating that many factors other than this are affecting the outcome.

### Table I. Gain in learning in five runs of Genetics and Evolution

<table>
<thead>
<tr>
<th>Term</th>
<th>( &lt;g&gt; ) ± S.D.</th>
<th>Enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 99</td>
<td>0.54 ± 0.32</td>
<td>13</td>
</tr>
<tr>
<td>Summer 99</td>
<td>0.87 ± 0.09</td>
<td>17</td>
</tr>
<tr>
<td>Fall 99</td>
<td>0.88 ± 0.21</td>
<td>24</td>
</tr>
<tr>
<td>Spring 00</td>
<td>0.54 ± 0.32</td>
<td>11</td>
</tr>
<tr>
<td>Fall 00</td>
<td>0.78 ± 0.21*</td>
<td>59</td>
</tr>
</tbody>
</table>

*- 3 students out of 59 did not take one of the pre or post-tests.
Conclusion

Apparently, automated testing offers an inexpensive, easily maintained objective tool for evaluating courses. Such a method is likely to be accepted by faculty who are comfortable with computers. This may well offer a good way to measure differences between courses with large numbers of students. On the other hand, it is clear that small classes offer advantages for learning. Unfortunately, the statistical significance of comparisons among small classes is low, unless one can accumulate results from many successive offerings of the class.

The WebCT software is easy to use, especially considering the benefits that it offers. Students seem to accept the testing method very well, based on the fact that I received no complaints about it. The uniformity of the grading, the speed with which the exams are graded, and the feedback that is often customized to the student’s answer, may be seen as better than “hand grading.” The WebCT software comes with a sample test that can be used for quickly training the students in the use of the testing tools. I used WebCT not only for pre- and post-testing, but also for quizzes and mid-term examinations. Thus, the students were quite familiar with the system by the time they took the post-test. For term examinations only, the students also submitted written responses to the questions. In grading these examination papers, I had only to look at the handwritten versions of answers that were scored as being incorrect by the computer. This saved time, and I was able to judge whether to assign partial credit, overriding the grade assigned by WebCT. This resulted in extra points for most students on most tests. Thus, the speed of computer grading was combined with the analytical judgment of the instructor. I believe this is a useful practice, which is judged as being fair by the students. Their experience with taking the tests this way is likely to have made it easier for the students to accept automated testing on the final examination. (Incidentally, by adopting a policy of not returning the originals of written exams, I avoided having to deal with a large number of appeals for re-grading of examinations. Students who needed a hard copy to prepare an appeal were provided a photocopy on request).

The pre- and post-tests differ in that the latter counts toward the student’s grade. I believe most faculty would think it unfair to include the pre-test in the students’ grades. Some may think it better to exclude the post-test score from the final grade, in order to keep students from stockpiling exams. (It is possible to print them). However, over time I have not noticed any improvement in the absolute scores on post-tests, suggesting that stockpiling has not been a problem. The use of electronic examinations minimizes the amount of paper circulating, and probably helps to maintain the integrity of the examination process. The other possible objection to including the post-test in the final grade is that the magnitude of the value of $g$, the gain in learning, may be exaggerated. However, as long as the methods are not changed from one semester to the next, the comparisons should remain valid. On balance, I think the policy of including the post-test in the final grade encourages students to remember more of the course material, and thus serves a useful purpose, with no serious disadvantages.
Literature Cited


Call for Applications

John Carlock Award

This Award was established to encourage biologists in the early stages of their professional careers to become involved with and excited by the profession of biology teaching. To this end, the Award provides partial support for graduate students in the field of Biology to attend the Fall Meeting of ACUBE.

Guidelines: The applicant must be actively pursuing graduate work in Biology. He/she must have the support of an active member of ACUBE. The Award will help defray the cost of attending the Fall meeting of ACUBE. The recipient of the Award will receive a certificate or plaque that will be presented at the annual banquet; and the Executive Secretary will provide the recipient with letters that might be useful in furthering her/his career in teaching. The recipient is expected to submit a brief report on how he/she benefited by attendance at the meeting. This report will be published in Bioscene.

Application: Applications, in the form of a letter, can be submitted anytime during the year. The application letter should include a statement indicating how attendance at the ACUBE meeting will further her/his professional growth and be accompanied by a letter of recommendation from a member of ACUBE. Send application information to:

Dr. William J. Brett, Department of Life Sciences, Indiana State University, Terre Haute, IN 47809; Voice -- (812) 237-2392 FAX (812) 237-4480; E-mail -- lsbrett@scifac.indstate.edu

If you wish to contribute to the John Carlock Award fund, please send check to: Dr. Pres Martin, Executive Secretary, ACUBE, Department of Biology, Hamline University, 1536 Hewitt Ave., St. Paul, MN 55104.

Call for Nominations

Honorary Life Award

The ACUBE Honorary Life Award is presented to ACUBE members who have made significant contributions and/or service to ACUBE and the advancement of the society's mission. The award is presented at the annual fall meeting of the society.

If you wish to nominate a member of ACUBE for this award, send a Letter of Nomination citing the accomplishments/contributions of the nominee and a Curriculum Vita of the nominee to the chair of the Honorary Life Award committee:

Dr. William J. Brett, Department of Life Sciences, Indiana State University
Terre Haute, IN 47809, Voice -- (812) 237-2392, FAX (812) 237-4480
E-mail -- lsbrett@scifac.indstate.edu