RELIABLE ASSESSMENT WITH CYBERTUTOR, A WEB-BASED HOMEWORK TUTOR

World Conference on E-learning in Corporate, Government, HealthCare, & Higher Education, E-Learn 2002, Montreal, Canada, October 15-19, 2002

David E. Pritchard[1] and Elsa-Sofia Morote
Massachusetts Institute of Technology
Cambridge 02139

Introduction

It is the end of semester and you have two students waiting to talk with you about their score in the final exam. You know what they want. Are you sure they deserve to fail? Are you wondering if that student who miraculously passed, deserved to pass? You are a fair teacher, and put all your effort into elaborating a fair final exam but have you ever tested how reliable your final exam was? Do you really believe that a 3-hour final exam will accurately reflect the precise "skill" level of every student?

Socratic electronic homework tutor, CyberTutor can integrate effectively instruction and assessment. The following study shows that CyberTutor assessment has about 62 times less variance due to random test error than a three-hour final examination, and 53 times less variance than twelve weekly tests. This huge improvement results from the combination of three factors. First, the typical student interacts with CyberTutor for at least 48 hours per term, 16 times longer than the 3-hour final exam and 10 times more than 12 weekly tests together, diminishing the effect of lucky guesses and careless errors. Secondly, by considering requests for hints, solutions, and importantly the number of wrong answers and the time the student takes to complete each part of the problem, CyberTutor can make a much better determination of the student's skill on a problem. Lastly if a problem is too difficult for the student as presented, CyberTutor's hints and spontaneous responses to wrong answers systematically adapt the difficulty of the problem to the student's skill level, and the CyberTutor assessment algorithm accurately determines that level.

Tutoring with Assessment

The recent advances on the Internet and World Wide Web have brought attention to the world of Web-based education. There is an increasing demand for use of computerized educational aids by students and faculty (Cao & Bengu, 2000). Teachers are not only using electronic homework systems but recently there has been growing interest in using tutoring agents in education systems (Boy, 1995). The positive effects on learning of these web-based electronic homework tutors have been confirmed by several studies such as Mestre et al. (2000); Ogilvie (2000); Thoennessen & Harrison (1996) and Pritchard & Morote (2002a). Some of these electronic systems also eliminate the task of grading, enabling teachers to shift efforts to instruction, but few offer an integral system of tutoring and assessment.

This paper demonstrates that an electronic tutoring program can collects data that enables a far more reliable assessment of students' skills than a standard examination. The following scenario illustrates the basic idea: Imagine that a rich ship-owner has hired Socrates to tutor his children. At the end of the month he desires to assess the amount they have learned.

[1] Inquiries about CyberTutor to Dr. David E. Pritchard, e-mail dpritch@mit.edu

a) Would you advise him to administer a standardized hour-long test to the children?

b) Ask Socrates how much they have learned?

The answer is straightforward. Clearly Socrates' many hours of individual interaction enables him to form a far more accurate and nuanced assessment of each student than could be provided by a one-hour test. Moreover, this integrated assessment is continuously updated and is used by Socrates to adjust the pace and content of the session. Importantly, this continuous integrated assessment is available without the stress and time loss of traditional testing.

The educational impact of real-time processing and reporting results of diagnostic electronic homework of results is recognized by Just in Time Teaching (Novak et al., 1999). It enables classroom instruction to be efficiently tuned to students' needs. The essential element is feedback between the web-based tutoring and classroom activities.

CyberTutor has a fine-grained assessment algorithm which permits the identification of skill deficiencies in foundational skills as each student moved through the course. This would allow tutors (or perhaps additional CyberTutor problems) to remedy the deficiency before it causes irreparable harm to the student. Additionally, knowing the class' skill profile on each detailed topic could allow the teacher's next lecture to review topics that were not being well learned.

Methodology

Using data of Introductory Newtonian Mechanics, course 8.01 teaching at spring 2001 at MIT we evaluated reliabilities of the final exam, weekly tests, and Socratic tutor. Over 90% of the students taking 8.01 in the spring term have previously attempted this course without being able to learn the problem solving skills demanded by the 8.01 examinations. Besides course elements such as class participation, written homework, group problem solving, and a Socratic electronic homework tutor, the students were given 12 weekly tests during the semester and a final exam.

We note in passing that this particular final exam had three features that should give it less error than typical physics finals. A standard test (Mechanics Baseline test) was used for 1/3 of the Final. The remainder was divided into basic skill problems and complex problems which accentuate the performance of different students, and the total test requested 47 specific responses from the students. In comparison, the 8.01 final in the preceding semester required only 28, and therefore probably had larger relative statistical fluctuations.

Testing reliability.

By definition, someone taking a test or other assessment measure achieves the "observed score." The observed score is thought to be composed of two different components: The true score, an exact measure of the amount of the trait or skill assessed, offset by a random error score (Observed score = True score + Error score). The error may result from individual student responses to particular questions (Manson & Bramle, 1997) or external conditions such as fatigue and noise.

The variance of the observed score is composed of the variance of true scores and the variance of error. Then, $\sigma^2_o = \sigma^2_t + \sigma^2_e$, where, σ^2_o = variance in observed scores, σ^2_t = variance in true scores, and σ^2_e= variance of error

The "reliability" of a test or assessment measures its repeatability or freedom from random testing errors. An assessment is considered reliable if it would give us the same result over and over again. By definition, the reliability is the ratio of variance of the true skill divided by the total variance of the measurement (observed variance), $\sigma^2_t \; / \; \sigma^2_o$. The reliability provides a direct guide to the reproducibility of the grade between two similarly prepared tests (odd or even numbered questions) on equivalent material. If reliability = 1, there is no error in the test.

A standard method of determining the reliability of an assessment is to divide the test questions into two subtests, for example the odd and even questions if they were equal in coverage of topics and difficulty. If the resulting subtests determine skill with perfect reliability, each student would receive the same score on both subtests, and a scatter plot of scores on the "odd (a)" and "even (b)" subtests would lie on the diagonal. We compute the reliability of the various assessments considered here from the *uneven split reliability formula* (see Pritchard & Morote, 2002a for derivations). This method does not assume that the halves have equal standard deviations, an assumption of the Spearman-Brown formula[2].

$$\text{Reliability} = 1- \; (1+ \beta)^2 / 4\beta * \sigma^2_{(a-b)} \; / \sigma^2_{(a+b).} + (1- \beta)^2 / 4\beta \qquad (1)$$

where $\beta = \__a \; / \; \__b$, the ratios of the means of the two subsets. In this formula, the key determiner of the reliability is the mean variation of the difference between the scores on the even and odd collections of problems, $\sigma^2_{(a-b)}$.

The Value of Electronic Homework Tutor as an Assessment Instrument

The reliability method was applied first to the MIT 8.01 final exam for 2001. Figure 1 (left panel) shows some lack of consistency between the performances on the two subtests in the Final exam. This reflects the error on the two subtests: they covered the same material, but random errors are considerable. Of course, adding the two halves together reduces the relative size of the error and results in reliability = 0.85 as computed from (1). In the same way, we computed the reliability of the 12 weekly tests resulting in a value of 0.87. The reliability of these tests, 0.85 and 0.87, is a moderate reliability in a test. Generally reliability above 0.90, typical of ETS[3] tests, is considered good for a test in education and behavioral sciences (Manson & Bramle, 1997). For example, in the case of the final exam, its reliability means that 85 percent of the variance in the observed scores is due to true score variance. The error causes 15 percent of the observed variance. This 15 percent error will offer lead to a misdetermination of who should pass and fail as we discuss later.

If someone is highly skillful, he generally make fewer errors and takes less time to successfully complete a task than a less skilful person. Moreover, any individual has the option to work more quickly at the expense of making more errors or viceversa. For the electronic tutor, we developed an assessment algorithm that depends on the number of right responses and hints, subparts, and number of solutions requested. We also included additional variables (time and wrong answers) as well as all other variables available from the log of each student's interactions with CyberTutor including among others, number of problems worked, etc. We call this assessment algorithm the "Enhanced CyberTutor Assessment"

The CyberTutor algorithm is computed from the even problems and the odd problems for each student (Figure 1, right panel). The curvature indicates a small excess of easier problems in the even group. The reliability from (1) of the combined assessment is 0.9976. This is a fantastic improvement over the final examination and weekly tests. That means that 99.8 percent of the observed variance in scores is due to true score variance. Overall, it has 62 times less error than final exam and 58 times less error than the all weekly tests together. It provides nearly error-free

[2] Spearman-Brown Reliability = 2 * [correlation (Odd, Even)] / [1 + correlation(Odd, Even)]

[3] Educational Testing Service, e.g. tests: GMAT, GRE, TOELF and PRAXIS

assessment, which as far as we know is an amazing record for high reliability for any assessment in education.

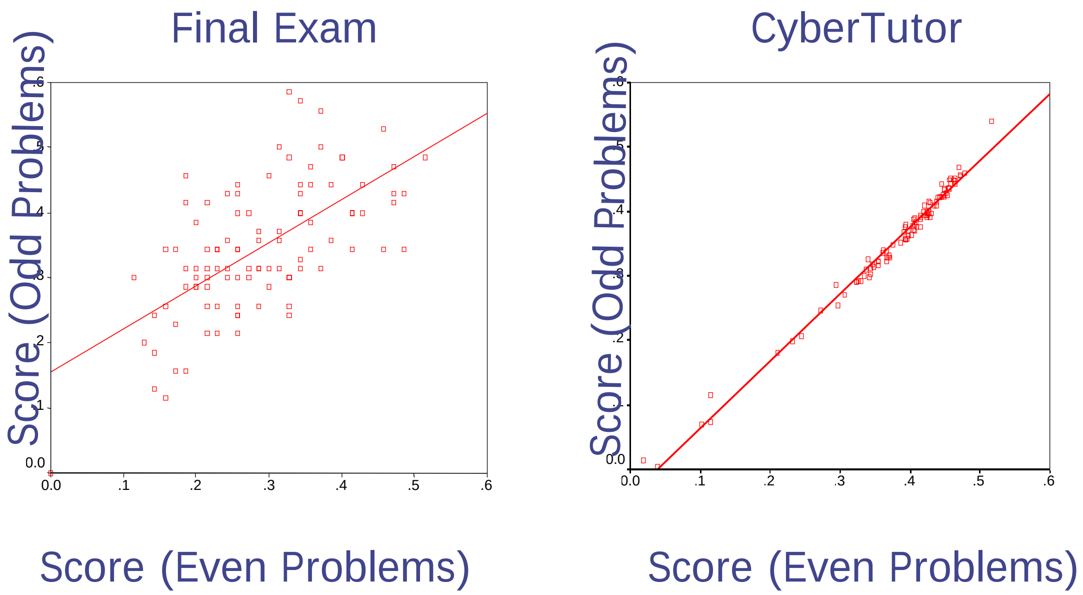


Figure 1. Grades of each individual student on the two subtests (odd -y axis- and even –x axis), comprising the CyberTutor and MIT 8.01 Final Exam in 2001.

Table 1 shows the differences among Final Exam, Weekly Tests and CyberTutor. The Enhanced CyberTutor algorithm shows higher reliability than the other tests. In addition, by eliminating marginally discriminating problems[4], we were able to increase reliability to 99.92 percent. "CyberTutor +" presents 183 times less variance than the final exam.

Table 1
Reliabilities' Summary

	Reliabilities	_ $^{2}{}_{e}$ /_$^{2}{}_{t}$	_e/_t
Final exam	0.85	0.17	0.42
Weekly tests	0.873	0.144	0.380
CyberTutor	0.9976	0.0024	0.0488
CyberTutor +	0.9992	0.0008	0.028

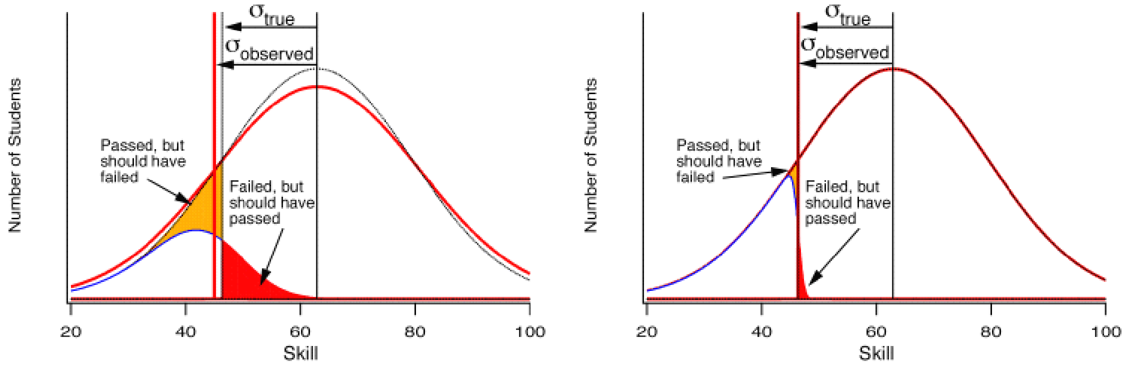
The final exam, which was carefully developed, contains so much error that 1/4 of the students assigned failing grades should has pass, and an equal number of students who should have failed are assigned a passing grade. Thus, for every two students who took the final exam one was misdiagnosed. A similar case occurs using weekly tests.

To see how this arises, look at Figure 2, where failure is set at one standard deviation below average, which fails 16 percent of the students, the typical failure rate for 8.01 course. The dashed line shows the true skill distribution inferred from the measured error and the

[4] By performing item analysis using item discrimination. It measures how well an item distinguishes between the students who understand the content universe if the test well and those who do not.

observed skill distribution (heavy line), which is slightly broader. The number of students failed versus their true skill level, is indicated by the small curve centered on skill = 40. (The width of this curve is very nearly the width of the error.)

Below skill = 35 the failed students of the true skill distribution, indicating that all students whose skill is this low fail. The scores between 30 and 46 (the selected true skill for passing) drop well below the true skill distribution indicating that some students (lighter shaded area) whose ability is below passing will pass. The distribution of failed students does not drop immediately to zero above true skill = 46, indicating that some students (darker shaded area) whose true ability is passing will fail the test.



Final Exam

CyberTutor

Figure 2. Fairness in the Final Exam versus CyberTutor

If the much more accurate Enhanced CyberTutor Assessment were used to set pass-fail instead of the final exam, only 3 percent of the students failed would be misdiagnosed (figure2, right panel). In a class of 100, only one student would be misdiagnosed, and his true skill would almost certainly lie only 1 or 2 points away from the pass-fail line.

Conclusion

We have shown that a web-based electronic homework tutor collects data which enables an assessment and tutoring and gives us a better way to deal fairly with students' capabilities. Teachers will stop wondering if a student who miraculously passed, deserved to pass, or if a student failed because of "bad luck". CyberTutor provides objective measurements on which educational decisions can be based. It can provide means for improving and enhancing feedback, motivation, and over learning.

The dramatic reduction in the variance of the error results from combination of three factors. First, the typical student interacts with CyberTutor for at least 48 hours per term, 16 times longer than the 3-hour final exam and 10 times more than 12 weekly tests together, diminishing the effect of lucky guesses and careless errors. Secondly, by considering requests for hints, solutions, and importantly the number of wrong answers and the time the student takes to complete each part of the problem, CyberTutor can make a much better determination of the student's skill on a problem. Lastly if a problem is too difficult for the student as presented, CyberTutor's hints and spontaneous responses to wrong answers systematically adapt the difficulty of the problem to the student's skill level, and the CyberTutor assessment algorithm accurately determines that level.

Tutor-based assessment has such low error that it is capable of providing accurate assessments on a number of pre-selected topics providing the detailed skill profile. If we took 20 topics – say the major topic of the 13 weeks of the course plus 7 foundational topics; the assessment on each topic would be twice times as reliable as the overall assessment of the current final. Even if we increased the skill resolution to include each of the major weekly topics (i.e. perhaps 180 topics) the variance on each would be small enough to identify those students requiring immediate assistance on a particular topic. By basing the assessment on a more careful analysis of wrong answers (was some particular step omitted or a particular wrong assumption made, or was this just a careless error?) we project that we can improve the CyberTutor assessment sufficiently to measure even this fine grid with the accuracy that the current final determines the overall skill of the students. CyberTutor merges assessment and tutoring and not only motivates students to learn but also helps them to learn and gives teachers much more accurate information to efficient and more closely tuned to students' needs.

References

Boy, G. (1995). Software agents for cooperative learning, in J.M. Bradshaw (ed.) AAAI Press/The MIT Press.

Cao, L., & Golden, B. (2000). Web-based agents for reengineering engineering education. J. Educational Computing Research, 23, 4, 421-430.

Ebel, R., Frisbie, D. (1991). Essentials of educational measurement. New Jersey: Prentice Hall.

Gregor M. Novak (Editor), Patterson, E., & Gavrin, A. (1999) Just-in-Time Teaching : Blending active learning with web technology. Prentice Hall Series in Educational Innovation.

Manson, E., & Bramle, W. (1997). Research in education. Sydney: Brow & Benchmark Publishers.

Mestre, J., Dufrense, R., Hart, D, & Rath, K. (2000). The effect of web-based homework on test performance in large enrollment introductory physics courses. Journal of Computers in Mathematics and Science Teaching

Ogilvie, C. (2000). Effectiveness of different course components in driving gains in conceptual understanding. Cambridge, Internal report, Department of Physics at MIT. [on-line] URL: http://torrseal.mit.edu/effedtech/

Pritchard, D. & Morote, E. (2002a) 99.9% Reliable Assessment with CyberTutor. Physics Department, Report to Massachusetts Institute of Technology. [on-line] URL: http://torrseal.mit.edu/effedtech/

Pritchard, D., & Morote, E. (2002b). What course elements correlate with improvement on tests in introductory Newtonian mechanics? NARST Annual Meeting New Orleans, Louisiana, April 7-10.

Thoennessen, M., & Harrison, M. (1996). Computer-assisted assignments in a large physics class. Computers Educ., 27, 2 , 141-147.