WHAT COURSE ELEMENTS CORRELATE WITH IMPROVEMENT ON TESTS IN INTRODUCTORY NEWTONIAN MECHANICS?[1]

Similar to National Association for Research in Science Teaching – NARST- 2002 Conference
New Orleans, April 7-10, 2002
Elsa-Sofia Morote and David E. Pritchard[2]
Physics Department
Massachusetts Institute of Technology
Cambridge, MA 02139-4307

We report the level of effectiveness of the various course elements (such as electronic homework, written homework, collaborative group problems and class participation) in calculus-based introductory Newtonian mechanics as taught to MIT students. Effectiveness is measured by the regression coefficient of a particular course element with the gain of the student's grade on the MIT final exam and on two widely used standard physics tests that emphasize conceptual knowledge: the Force Concept Inventory and the Mechanics Baseline tests. We find that the electronic homework as administered by CyberTutor is the *only* course element that contributes significantly to improvement on the final exam, and that CyberTutor and collaborative group problem solving contribute most strongly to improvements on the standard conceptual tests. We also report surveys that demonstrate strongly increasing student assessment of CyberTutor over the four terms of its use.

Theoretical Underpinnings

Physics education research has developed both in terms of the knowledge of teaching and learning, and curriculum projects and practices. Van Aalst (2000) examines the advances on how curriculum innovations have made an impact on physics learning. Tools to improve class teaching such as *interactive lecture demonstrations* (McDermott and Trowbridge, 1980) and instructional techniques such as *peer instruction* (Crouch and Mazur, 2001) and *group problems* (Heller, Anderson, & Keith, 1992) have been designed to increase conceptual learning as typical measured using instruments such as the Force Concept Inventory (FCI) and Mechanics Baseline (MB) tests.

Hake (1998) has conducted an analysis of pre- and post-test data obtained from the FCI, and the MB test to compare *interactive-engagement* versus traditional methods. Both tests are complementary probes for understanding of the most basic Newtonian concepts. Questions on the FCI test (see Hestenes, Wells, & Swackhammer, 1995). were designed to be meaningful to students without formal training in mechanics and to elicit their preconceptions about the subject; in contrast the MB test emphasizes concepts that cannot be grasped without formal knowledge of mechanics (Hestenes & Wells, 1992). Hake obtained data from both tests administered to 6,500 students in 62 courses, using the normalized, the improvement in score normalized by the maximum possible improvement as a metric for "course effectiveness in promoting conceptual understanding." The normalized gain is determined from the "after" and "before" examination scores (S):

[1] This work was supported by NSF grant PHY-9988732
[2] Inquiries about CyberTutor to Dr. David Pritchard (dpritch@mit.edu)

$$Normalized\_Gain: NG = \frac{S_{after} - S_{before}}{100\% - S_{before}} \quad (1)$$

Hake found that classes that used interactive-engagement methods outperformed traditional classes by almost two standard deviations with respected to the normalized gain. He found that traditional classes have an average normalized gain equal to 0.23 whereas classes using interactive methods obtained an average gain of 0.48 ±0.14 (std dev). In the same way, utilizing the FCI test, Jeff Saul (1998) compared student learning of mechanics in traditional (lecture and recitation) first-semester calculus-based physics with three innovative curricula: McDermott's *tutorials,* Heller's *group problem solving,* and Law's *workshop physics* (lecture, lab and recitation combined into three two-hour guided discovery lab sessions per week). As in Hake's study Saul confirmed that traditional classes average about 0.20 normalized gains, and the innovative curricula (tutorials and group problem solving) average 0.37 gains, while guided-discovery instruction (workshop physics) averaged 0.43 of normalized FCI gain.

Craig Ogilvie (2000) used a similar method but added an important course element not present in Saul's analysis: electronic homework. Applying the FCI test before and after the course in one classroom of approximately 100 students at Massachusetts Institute of Technology (MIT), Ogilvie gave data on the effectiveness of the various course elements such as *tutorial attendance*, *written problem sets*, Pritchard's *electronic homework tutor* (CyberTutor, 2001) and group problem solving. He concluded that electronic homework tutoring led other course elements in producing gains in FCI score that were twice as large as those from the written problem set. Solving problems in groups led to intermediate gains in FCI score.

Homework in general has been appreciated as an important course element. For instance, Cooper (1989) found at least 50 studies that correlated the amount of time students reported spending on homework with their achievements. Cooper affirms that homework has several positive effects in achievement and learning such as better retention of actual knowledge, increased understanding, better critical thinking, and curriculum enrichment.

Electronic homework as a course element has even more positive effects than written homework affirm some researchers (Mestre et al., 2000; Ogilvie, 2000; Thoennessen and Harrison, 1996). Mestre et al. (2000), for example, compared the effect of the electronic homework and the written homework on student achievement as measured by exam performance. They found that electronic homework led to higher overall exam performance. Thoennessen and Harrison also affirm that electronic homework has a clear correlation with the final exam score, and students prefer using it over written homework.

In this present study, we analyzed the effects of course elements in the MB test gain and the MIT final exam gain on the introductory Newtonian mechanics course at MIT.

Course Overview

Calculus-based introductory Newtonian mechanics, course 8.01 at MIT, is the most difficult required course for entering freshmen. Typically 15% of the students fail to receive a grade of C or better and hence are forced to repeat it. Consequently 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 – mostly quantitative problems with symbolic answers. The spring course has been reorganized to teach these problem solving skills.

The revamped course, in which these studies were correlated, does not use lectures as vehicles for presentation of new material – not a radical step since these students have had an opportunity (but not a requirement) to attend 8.01 lectures with demonstrations previously. “New” material is introduced in three recitations on Monday through Wednesday, reviewed in tutorials on Thursday, and reviewed and tested on Friday. Homework problems were presented two ways: in conventional written form and electronically, using CyberTutor. Attendance and participation in recitations constitutes 3% of the grade, and a challenging group problem was given to collaborations of two or three students in class each week which counts as 7% of the grade. Small tutorials (three or four students) were required of all students in 2000, but were required only of under-performing students in 2001.

Course Elements

The spring course consists of the following course elements:

CyberTutor. CyberTutor behaves like a Socratic tutor, offering students help upon request in the form of hints and simpler subproblems, spontaneous warnings and helpful suggestions when wrong answers are given. It leads 90% of the students to the correct solution, so the CyberTutor grade is primarily an indication of how many problems were attempted with it.

Written Homework. Solutions were provided on the due date, and all the problems were subsequently graded. The grade is strongly correlated with the amount done.

Group Problems. Students worked in groups of three to collaboratively solve complex problems as pioneered at the University of Minnesota (Anderson, Heller and Keith, 1992.)

Class Participation. (2001 only). Participation is graded based on attendance at and participation in recitations in a 2:1 ratio. (There were three recitations/week in this course; and only one lecture.)

Tutorials. Three or four students met with a senior undergraduate or graduate tutor for a one-hour tutorial. (2000 only; required only for less skillful students in 2001).

Methodology

The Force Concept Inventory was administered before and after the 8.01 course in Spring 2000 by the professor in charge C. Ogilvie (2000). The MB test is more balanced between concepts and numerical problems than the FCI test and also includes energy and momentum, was administered before and after the Spring 2001 course by Prof. Pritchard. In addition, the *gain* on the final examination was computed for those students who had taken one of final exams in the 8.01xx sections the prior semester. The normalized gain in all three cases was found using formula (1).

This study uses the standard statistical techniques of regression and multiregression. These infer the performance gain associated with use of the various course elements by studying the difference in gain of students who use a particular course element more or less. Multiregression provides a valid statistical method for isolating the effects of single course element on the normalized gain.

It is worth noting the differences between our methodology and the more common one of giving class A one treatment and class B another. This type of study is ideal for deciding which treatment is better, but determines only the differential effect of the treatments. In contrast, multi-regression shows the effects of a single course element. Moreover, it can compare several different course elements in one study, whereas the A vs B approach becomes much more difficult when more than two factors are being compared. One drawback of our statistical approach is that it requires a larger sample to produce results of the same statistical validity as the A vs B approach. We note, however, that the prime cause of the scatter in our data (e.g. in Fig. 1) is the random error in the assessments used and the fact that this error adds when subtracting scores to compute the gain.

Gain on the MIT Final Exam

The majority (72%) of the students in 8.01 in the spring of 2001 had taken a final examination in one of the four 8.01xx courses the preceding semester. This gave us a "before" and "after" final examination grade from which to calculate the normalized gain on the final (NGF). To find which course elements correlated with the gain, we first plotted the gain of the final versus each of the course elements, considered independently. For each scatter plot we fit a straight line using a linear regression. Scatter plots for CyberTutor and written homework are shown in Figure 1.

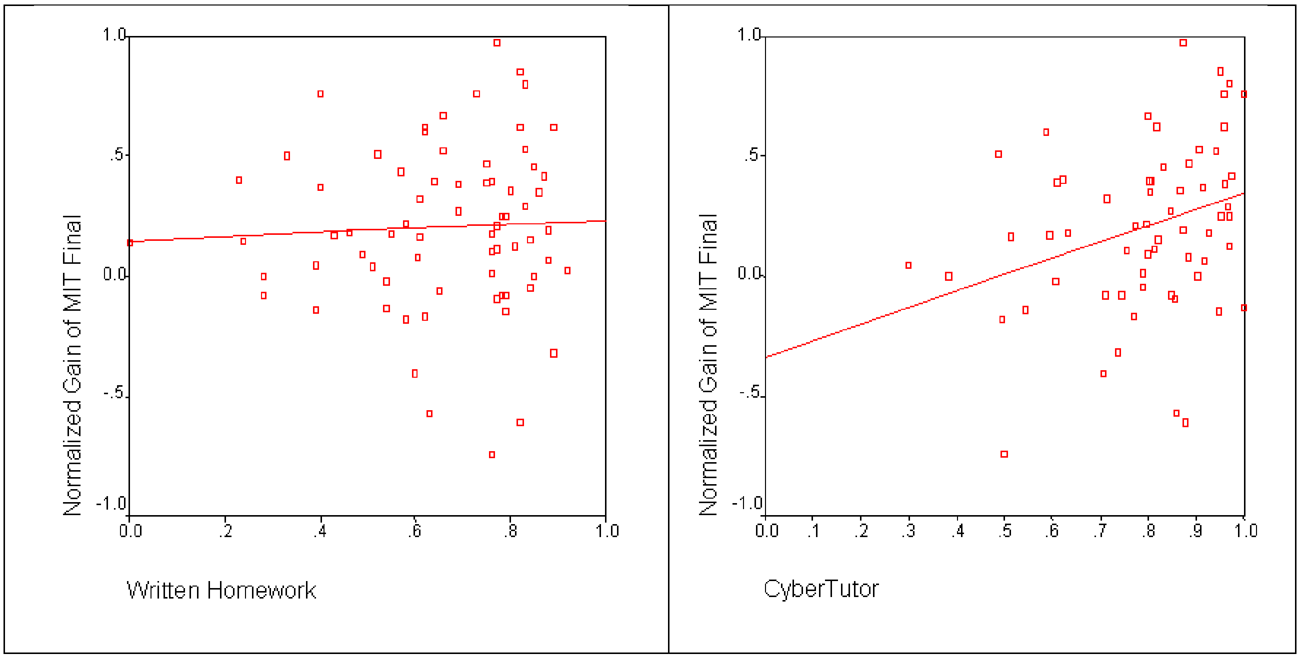


Figure 1. Gain on Final versus Written Homework (Left Panel) and vs. CyberTutor (Right Panel).

In Figure 1, straight lines show fits using linear regression. The CyberTutor slope indicates a gain of 0.55 for the average student relative to the intercept CyberTutor = 0. The slope of written homework is small and statistically insignificant. Written homework is often done in isolation, and the student focus is simply getting the answer (Johnson, 2001). However, CyberTutor offers follow-up comments and follow-up questions to make students ponder the significance of their correct answers. A second factor is that copying of written homework is endemic and has low instructional value. In contrast, study of the student response patterns on CyberTutor showed only two students whose lack of wrong answers and hints requested strongly suggest that they were obtaining the solutions elsewhere.

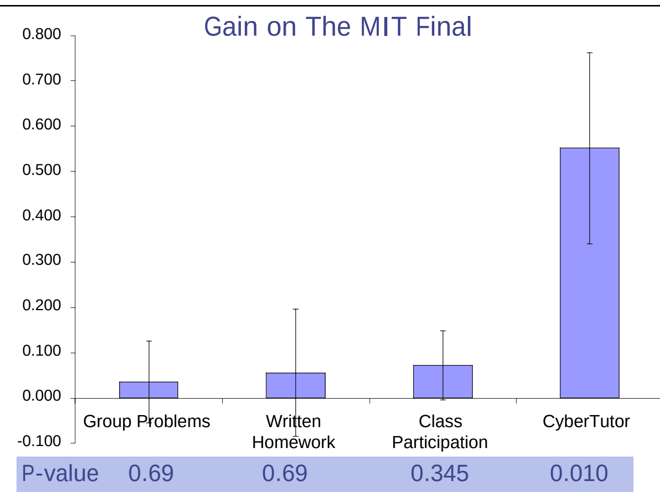
Data from plots including those on Figure 1 are summarized in Table 1 which shows the slope (β coefficient in first column). In the next two columns, we calculate the gain attributable to each element by multiplying its slope times the average score on that course element, along with its standard error (δ gain), which is the standard error in β times the average score. The last column represents the p-value, the probability that the observed value of β could result from chance alone (McCall, 1998). CyberTutor has the highest slope (0.688) and a very significant (p-value[3] =0.01). Written homework, group problems and class participation show no significant contribution to the NGF (note the difference on the slopes of written homework and CyberTutor in Figure 1). These results were confirmed by treating these data using a multiple linear regression, which confirmed that CyberTutor is the only significant course element associated with gain (0.551) on the final exam. The mutiregression analysis shows that CyberTutor contributed 72 % of the observed gain on the final exam and the other course elements the reminder 28%.

The gain inferred for CyberTutor 0.551 ± 0.21 (Table 1) is a remarkable gain for a single course element - an entire course is considered good if it yields a gain greater than 0.4 on tests like the MB and FCI tests which are narrowed in focus than the MIT final. The error in the gain _ gain = SD(_)*mean, is shown within error on the bar graph (Right panel).

Table 1.

Results for Linear Regression Fit for Gain on The Final Exam vs. each Course Element Considered Independently

Course Element	Slope (β)	Gain = β*mean	δ gain	p-value
CyberTutor	0.688	**0.551**	0.211	**0.010**
Written Hk.	0.083	0.055	0.140	0.690
Group Problems	0.056	0.035	0.090	0.690
Class Participation	0.116	0.072	0.075	0.345



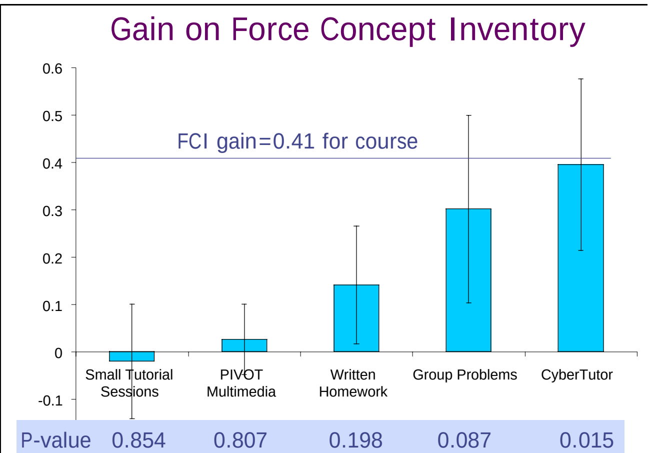
[3] A p-value of less than 0.1 leads to rejection of hypothesis of no regression

The Gain on Force Concept Inventory and Mechanics Baseline Tests

The Force Concept Inventory test was administered before and after the course taught by Prof. Ogilvie in Spring 2000. Correlated the gain with various course elements. Scatter plots of these data versus course elements are contained in Ogilvie (2000). Based on individual regressions, it was found slopes, gains, δ gains and p-values[4] for each element (Table 2). CyberTutor and Group problems contributed most significantly to the FCI gain.

Table 2
Gain on the Force Concept Inventory for each Course Element

Course Elements	Slope	Gain	δ gain	p-value
CyberTutor	3.73	0.395	0.181	**0.02**
Written Homework	1.66	0.141	0.124	0.2
Group Problems	3.87	0.301	0.198	0.09
Small Tutorial Sessions	-0.27	-0.02	0.121	0.85



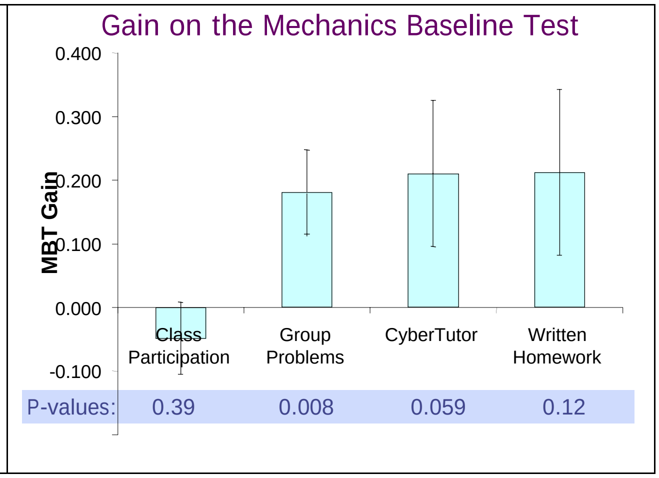
Based on Ogilvie (2000).

Professor David Pritchard administered the Mechanics Baseline test. Table 3 shows individual regressions between each course element and MB normalized gain. Group problems and CyberTutor contributed most significantly to improvement on the MB test (p-value < 0.1). Written homework showed higher gain than group problems, but it has high p-value. Class participation had no significant effect. By applying multi-regression, due correlation between variables, the only statistically significant variables (p-value < 0.05) identified were CyberTutor (39% of the observed gain) and group problems (46%), which together contributed 85 % of the MB normalized gain. These results could be affected at the gain 0.06 level by incorrect grading of one of the problems, which was discovered only after the tests were destroyed.

[4] Gain, δ gain and p-values were evaluated based on data in Ogilvie (2000).

Table 3
Gain on Mechanics Baseline Test for each Course Element

Course element	Slope	Gain	_ Gain	p-value
CyberTutor	0.264	0.21	0.125	**0.059**
Written Homework	0.297	0.212	0.131	0.109
Group Problems	0.267	0.181	0.066	0.008
Class Participation	-0.072	-0.049	0.057	0.39



In summary, significant contributions to the gain on these more conceptual tests (MB and FCI) come from both CyberTutor and group problems. It is encouraging that CyberTutor, not designed to teach concepts, compares well with a technique known to teach concepts effectively.

Student Opinion about CyberTutor

We now discuss student opinion concerning the educational effectiveness of CyberTutor and the desirability of using it in the future. This provides complementary information about CyberTutor's effectiveness, and about its overall level of student acceptance. The significant term-by-term increase of both factors indicates the desirability of using it in future classes.

We have asked two questions fairly regularly on the end of term questionnaires about CyberTutor, one to assess learning relative to written homework, and the other to address whether students thought it should continue to be used. The strong upward trend of the data on the accompanying graphs indicates that continued use of CyberTutor is strongly recommended, most recently by a 5:1 ratio (Figure 3, right panel). The underlying cause for this recommendation may well be that the students feel that they learn significantly more per unit time when using CyberTutor instead of doing written homework (Figure 3, left panel). We confirm studies of Thoennessen and Harrison (1996) that students prefer electronic homework over written homework.

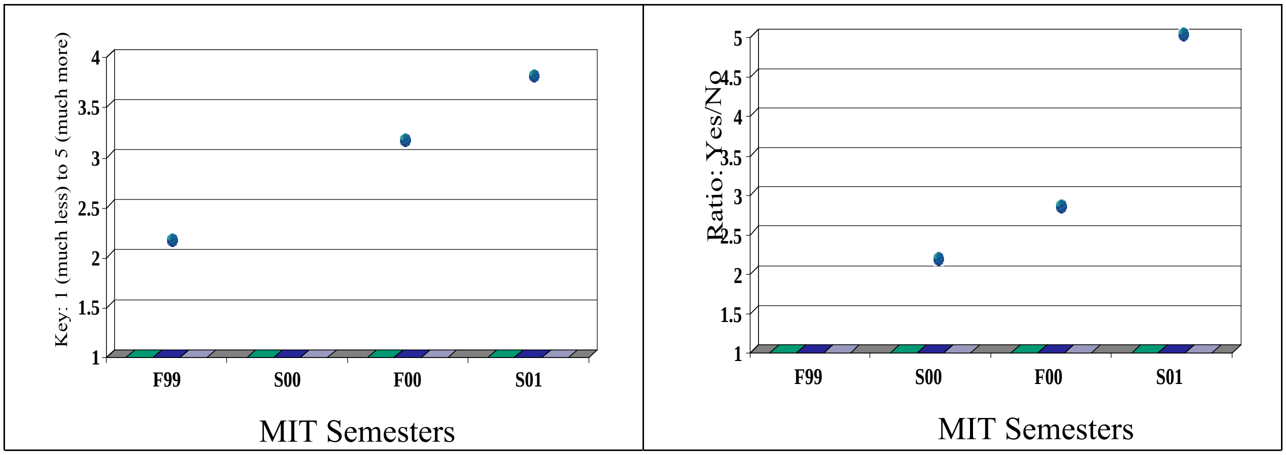


Figure 3. Right panel: Average student response to "compare the amount you learn per unit time using CyberTutor with time spent (including studying the solutions) on written homework." Left panel: Ratio yes to no students' responses to the question "Would you recommend CyberTutor for use in the 8.01 next year?"

Discussion

Based the three independent studies (derived from contributions to MIT final, MB test, and FCI test gains), we can say with statistical assurance (product of values p-values 0.00001) that students who elect to do more CyberTutor homework significantly improve their scores on the assessment instruments relative to their performance on these instruments before using CyberTutor. Furthermore, we can state with assurance at the $p \approx 0.01$ level that it helps on the final exam and that it is the course element that outperforms all the others on the combined before and after tests.

It is tempting to dismiss this as "just a correlation – perhaps the better students found CyberTutor easier to use and used it more". Such simple arguments fail, however, when one realizes that the correlations documented here are with *gain* not with *score*. The better students would have done better on the before test as well as the after test, hence no correlation with being a better student will influence the gain. There is a strong inference that students learn the material on the tests by using CyberTutor.

The average normalized gain has become a traditional figure of merit for measuring the increase in conceptual understanding occurring in a particular course. For that reason we use the normalized gain as the dependent variable, and have defined effectiveness in terms of the gain associated with each course element. In order to justify this definition, we must discuss whether the course elements represent learning activities, or simply provide assessment of material learned elsewhere.

Obviously the recitation participation grade and tutorial attendance are almost purely instructional as the scores indicate only that the students partook of these activities, not that they did so skillfully. The grades on CyberTutor and Written Homework are primarily an indication of the number of problems and assignments attempted, and are therefore largely instructional. This is particularly true of CyberTutor where about 90% of the students starting any problem part obtained the correct answer (and received essentially full credit since less than one hint is taken per part, and it costs only a 3% penalty; wrong answers were not penalized.) The group problem is obviously partly an instructional element and partly an assessment. At the purely assessment end of the scale are the weekly tests. The fact that these exam grades correlate with normalized gain of the final does not indicate that taking these tests is instructional - only that the material had already been learned elsewhere. Moreover, one can hardly advise a weak student to spend more time taking group problems or weekly tests even if they do have some instructional value. For these reasons weekly tests were not included in this study.

Clearly recitations and tutorials are purely instructional elements. Unfortunately they don't help: attending tutorials was shown to have no correlation with improvement on the FCI test, and attending and participating in recitation had no correlation with improvement on either the MB test or on the MIT final exam. In fact two of the three regression coefficients were negative (although insignificantly so), so this study can quite certainly state that attendance at class and tutorials definitely does not help improve test scores. This is not surprising to those who know of the general ineffectiveness of passive learning activities, but it will come as a disappointment to professors who work hard to make their recitations effective. Moreover, if these data were widely disseminated among students, the result would hardly be an increase in class attendance. The only ray of hope for believers in the efficacy of personalized instruction is that the students in spring term 8.01 might be there because they do not learn in a classroom setting.

Conclusion

This study shows that CyberTutor has been the most effective course element of spring term 8.01 over the past two years. It is the only instructional element of the course that contributes significantly to gain on the final exam, and it contributes strongly to the gain on the conceptual tests (MB and FCI). The group problem is the second most effective element by virtue of its effectiveness on the standard tests emphasizing conceptual knowledge. CyberTutor also receives an improving and now very strong recommendation from the students that it be used in the future.

On the other hand, attending tutorials was shown to have no contribution with improving on the FCI test, and attending and participating in recitation had no contribution with improvement on either the MB test or on the MIT final exam. This suggests that efforts to improve the instruction should concentrate on improving CyberTutor and on finding recitation and tutorial formats that are more effective instructionally. Recent educational research offers some suggestions for improving instructional formats.

References

Anderson, A., Heller, P., & Keith, R. (1992). Context-rich problems. Am. J. Phys 60, 627.

Cooper, H. (1989, November). Synthesis of Research on Homework. Educational Leadership, 47, 3, 85-91.

Crouch, C., & Manzur, E. (2001, September). Peer Instruction: Ten years of experience and results. Am. J. Phys, 69, 970-977.

Hake, R. (1998). Interactive engagement versus traditional methods: A six thousand student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66, 64.

Hestenes, D., & Wells, M. (1992, March). A mechanics baseline test. The Physics Teacher, 30, 159-166.

Hestenes, D., Wells, M., & Swackhammer, G. (1995). Revised 1995 FCI version Phys. Teach, 30, 3, 141.

Johnson, M. (2001, July). Facilitating high quality student practice in introductory physics. Phys. Educ. Res., Am. J. Phys. Suppl., 69, 7.

Lansford, C. (1999). Using Pre-test/Post-test data to evaluate effectiveness of computer aided instruction. University of Texas at Brownsville. ERIC ED432 2663

Lee, F., & Heyworth, R. (2000). Electronic Homework. J. Educational computing research, 22, 2 171-186.

McCall, R. (1998). Fundamental statistics for behavioral sciences. NY: Brooks/Cole Publishing Company.

McDermott, L. & Trowbridge, D. (1980). Am. J. Phys. 48, 12, 1020.

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. (2001). CyberTutor, an electronic homework tutor [on-line] URL: http://torrseal.mit.edu/effedtech/

Saul, J. (1998). Beyond problem solving, evaluating introductory physics courses through the hidden curriculum. Ph. D. Dissertation, University of Maryland.

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

Van Aalst, J. (2000). An introduction to physics education research. Can J. Phys. 78, 57-71.