Report on Visual Physics Program for Non-Physics Majors in Fall 2003 Semester

P. McIntyre, J.-H. Byeon, C. Ezrailson, T. Kamon, V.E. Mayes, C. Raisor, A.V. Sokolov, D. Toback, and A. Woodward
Department of Physics, Texas A&M University

N. Simpson
Center for Teaching Excellence, Texas A&M University

M. E. Troy
Measurement of Research Services, Texas A&M University

June 3, 2004

Abstract
Visual Physics (VP) is a systematic reform of the lab and recitation sessions in PHYS-218, the calculus-based freshman physics course. This report presents the results from the first-year trial program for non-physics majors and assessments of its effectiveness in teaching physics and technical writing, compared to sections taught using the traditional approach. The exam scores, along with technical writing scores and force concept inventory test scores, for the students in the VP lab are compared to those in the traditional lab in the fall 2003 semester. Although the statistics are limited, the data indicate that the VP lab with a rigorous technical writing component improves students’ exam scores (and, therefore, their letter grades in the course), and their ability to write a scientific report.
I. Introduction: Program Rationale and Research Questions

For many years, lab and recitation in the first-year, calculus-based physics course (PHYS-218) for non-physics majors at Texas A&M University (TAMU) have been led by graduate teaching assistants (TAs) with little faculty supervision and without any clear connection to the lectures. There was no rigorous grading of students’ lab reports (the score of each lab report was usually better than 90!), and TAs offered no feedback to students on how well they communicated what they were learning during the semester.

Several attempts have been made to improve the quality of feedback students receive on homework. Of them, an online homework grading system, CHAGS, in WebCT [1] has been most successful in helping reduce the number of D’s and F’s students receive in PHYS-218 and in improving their grades. However, students still relied on memorizing problem-solving steps instead of knowing how to analyze critically, which involves applying thinking skills beyond memorization and comprehension, skills such as analysis and synthesis. Further, the early program did not address students’ different learning styles. Professors’ lectures and corresponding labs incorporated narrowly-defined teaching strategies, which missed students having different learning styles.

A visual physics (VP) 218 program at TAMU is a fresh approach to recitations and laboratory experiments that is accompanied by technical writing (TW) instruction in PHYS-218. The approach was inspired by the work of Pat and Ken Heller at the University of Minnesota [2] and decades of research in how students learn first-year physics [3]. In keeping with the Minnesota model, a goal of the VP program, developed by Dr. Peter McIntyre, was to create a learning experience in lab, recitation and technical writing sessions that builds students’ understanding and confidence in using the tools of physics and in communicating what they have learned. Thus, they would succeed in the lecture exams and emerge with a mastery that can support them in subsequent courses.

This report describes the key components of the VP program, followed by an evaluation of the performance of the VP program with the following eight questions:

1. Are there any pre-existing differences in students’ academic backgrounds prior to the fall 2003 semester that would affect their performances in VP?

2. Are there any pre-existing differences in students’ understanding of basic concepts in Newtonian mechanics prior to the fall 2003 semester?

3. Does the VP technical writing component ask too much from students?

4. How does the VP program help the students’ final letter grades?

5. Does the VP program improve exam scores?

6. What is the impact of TW instruction on writing the scientific lab report?

7. Were the interactive-engagement methods effective? If so, how?

8. Does the VP program improve students’ conceptual understanding?

---

*a An ideal reform would also have included a reform of lecture to maximize students’ mastery of the physics.*
II. Visual Physics: Program Description

The following outlines key components of the program (also see Appendix A).

(1) **Visual:** Each lab consists of an experiment with a video camera and a PC to allow students to convert the visual data of the object’s motion into a numerical data and into a graph (on the PC). These visuals are also used to communicate via a lab report the results of the experiment.

(2) **Proficiency in real software tools:** The software used in the course is LabView™, Microsoft EXCEL®, Microsoft WORD® and AutoCad™ programs. These are professional tools that the students will use throughout their college education and then in their careers as either scientists or engineers. We are equipping them in the first semester as they use these tools to complete each experiment and corresponding lab report.

(3) **Small Size:** Each lab is limited to 21 (24 in the spring 2004 semester) students, compared to 32 in traditional (TR) lab.

(4) **Team Work:** During recitation, TAs review homework problems with students. Teams of three students work together in cooperative groups to synthesize solutions to complex problem scenario quizzes. The TA intervenes to guide students toward the next step of solution, and to answer general questions.

(5) **TA Instruction:** A week-long intensive training session is provided prior to the start of the current semester as well as weekly training for recitation and lab cooperative group and pedagogical instruction. The training and instructional sessions are led by Cathy Ezrailson, a physics education researcher in the College of Education.

(6) **Writing a scientific paper for each lab:** Students receive technical writing instruction each week, taught by Cindy Raisor, on how to write a serious 2-page scientific paper, or lab report, that demonstrates their understanding of the lab experiment. We have adopted a particular format/style, IEEE, for the students to develop this report. Using the IEEE format limits the number of pages of the report so that students must write concisely. Further, each lab report is graded based on a rigorous grading rubric to give feedback to students in each of the major areas of content and style.

Figure 1 illustrates the different styles and the VP components. In the traditional format, students often see a route of [Lecture] $\Rightarrow$ [Homework practice] $\Rightarrow$ [Exam] in their leaning path. Three contact hours a week in recitation and lab were insufficiently interacting with students. Therefore, our goal for the new approach was to use more efficiently the three-hour period in order to improve students’ mastery of the subjects.

In the new approach the “visual” and “team work” components (see Figure 2 and Figure 3) reinforce what the students learn in the same week (or one week later) as the lecture is given. The “writing” component requires students to synthesize what they know about physics, which also reinforces the problem-solving-skills they are applying in the other components. **One might want to ask why we include a serious writing component with Visual Physics?** Our approach to the course is to use writing as a tool for both communicating AND learning scientific information. Weekly sessions merge two goals of writing instruction:
(i) learning to write a specific product most commonly associated with the academic discipline (in this case, the 2-page scientific paper);

(ii) writing to learn a subject (in this case physics).

Therefore, instruction emphasizes the process of writing as much as the product, equipping the students with skills they will apply in their current and future coursework, such as PHYS-208 (electricity and magnetism), as well as in their professional careers. These goals reinforce the university’s commitment to using writing as a learning tool, one of the goals of the Writing-Intensive courses: “Writing to learn ‘content’ and the way members of the discipline think, argue, or research is as important as writing to learn a discipline’s ways of presenting material” [4].

Figure 1: Sketch of learning styles. Taken from workshop handouts of “Effective Teaching,” Richard M. Felder (N.C. State Univ.), Texas A&M University, January 30-31, 1998. Connection to PHYS-218 is also illustrated. The number of hours in each parenthesis is the contact hours a week. Each component of the VP program is shown in shaded (green) callout, appealing to learning style(s). “Visual lab” appeals to visual learners, “cooperative recitation” enhances students’ active process, while “writing a scientific lab report” along with “technical writing” may appeal to organization and understanding. The impact of “lecture” depends on each instructor’s teaching style. “LabView™, EXCEL®, and WORD®” are skills that students will use throughout their college education.
Visual Physics

Figure 2: Selected photos from the fall 2003 semester showing small class size and team work in recitation where V.E. Mayes (left), J.-H. Beyon (middle), and A. Woodward (right), Visual Physics TAs, were intervening with students during recitation.

Figure 3: Selected photos from the fall 2003 semester showing students who were working together in setting up an experiment (left) and a student reviewing a motion picture before the analysis (right) in video-based “visual” laboratory experiment.
III. Evaluation: Student Performance and Overall Program

The VP program began with three honors sections in the spring 2003 semester. In the fall 2003 semester, Dr. McIntyre taught three sections of honors and expanded the scope of implementation to six sections of the regular (non-honors) and prepared for a serious assessment of the new strategy in comparison to the traditional strategy. Three professors (A, B, and C) participated in the VP program. Each had two sections with VP lab and two sections with TR lab. Students in VP sections, who were re-taking PHYS-218, were removed from this analysis because they were excused from taking the lab and did not receive any benefit from VP lab. For consistency, re-taking students in TR sessions were also removed. Table 1 is a summary of the VP program.

Table 1: Summary of visual physics program in the fall 2003 semester. The number of students does not include re-taking, Q-drop/withdraw and incomplete students. There are two students who withdrew in Dr. A’s VP group, whom we don’t count for this study.

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Total Number of Students</th>
<th>VP lab with TW</th>
<th>VP lab without TW</th>
<th>Traditional lab without TW</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>86</td>
<td>-</td>
<td>34 (group VP0)</td>
<td>52 (group TR0)</td>
</tr>
<tr>
<td>B</td>
<td>78</td>
<td>32 (group VP1)</td>
<td>-</td>
<td>46 (group TR1)</td>
</tr>
<tr>
<td>C</td>
<td>56</td>
<td>26 (group VP2)</td>
<td>-</td>
<td>30 (group TR2)</td>
</tr>
</tbody>
</table>

Several assessment tools were administered at the beginning of the semester and were done again at the end of semester. Student background information was collected to analyze the population of each group. The performance of course exams was tracked to evaluate the effectiveness of VP in enhancing students’ learning of the physics. These results are summarized below.

1. Are there any pre-existing differences in students’ academic backgrounds prior to the fall 2003 semester that would affect their performances in VP?

For an assessment such as this, it is key to prepare three VP groups of non-honors students whose academic backgrounds are identical. We measured their academic preparation before taking PHYS 218 course by looking at their high school (HS) class rankings and their SAT math and verbal (SATM and SATV) and ACT verbal (ACTV) scores. Those are summarized in Table 2.

The VP0 group has higher scores in all of SATM, SATV, ACTV and HS ranking categories. We immediately notice that none of the pair-wise comparisons would be significant from the Math ANOVA analysis. Scheffe Post Hoc tests [5] show that the differences among the three groups on SATV, ACTV and HS ranking scores are not significant except ACTV scores between VP0 and VP1/VP2 whose significance is obtained to be at 0.016. We conclude that homogeneity of three groups are acceptable for our assessments.

The mean difference is significant if the significance is below 0.05.
Table 2: Summary of SAT math/verbal and ACT verbal scores and high school (HS) rankings. The value in each parenthesis is the significance by Scheffe Post Hoc tests [5] comparing to the VP0 group. Note that the numbers of students do not match with those in Table 1. This is because we used an early version of the student list. However, this does not change our conclusion on the homogeneity of three groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR0/TR1/TR2</td>
<td>SAT Math</td>
<td>122</td>
<td>460</td>
<td>800</td>
<td>659.8 (n/a)</td>
</tr>
<tr>
<td></td>
<td>SAT Verbal</td>
<td>122</td>
<td>370</td>
<td>800</td>
<td>610.1 (0.656)</td>
</tr>
<tr>
<td></td>
<td>ACT Verbal</td>
<td>58</td>
<td>15</td>
<td>36</td>
<td>26.5 (0.107)</td>
</tr>
<tr>
<td></td>
<td>HS Percentile</td>
<td>129</td>
<td>3</td>
<td>100</td>
<td>87.4 (0.741)</td>
</tr>
<tr>
<td>VP0</td>
<td>SAT Math</td>
<td>33</td>
<td>550</td>
<td>800</td>
<td>681.5</td>
</tr>
<tr>
<td></td>
<td>SAT Verbal</td>
<td>33</td>
<td>510</td>
<td>800</td>
<td>623.6</td>
</tr>
<tr>
<td></td>
<td>ACT Verbal</td>
<td>14</td>
<td>24</td>
<td>35</td>
<td>29.1</td>
</tr>
<tr>
<td></td>
<td>HS Percentile</td>
<td>33</td>
<td>39</td>
<td>100</td>
<td>89.7</td>
</tr>
<tr>
<td>VP1/VP2</td>
<td>SAT Math</td>
<td>65</td>
<td>440</td>
<td>800</td>
<td>654.8 (n/a)</td>
</tr>
<tr>
<td></td>
<td>SAT Verbal</td>
<td>65</td>
<td>340</td>
<td>790</td>
<td>591.1 (0.106)</td>
</tr>
<tr>
<td></td>
<td>ACT Verbal</td>
<td>30</td>
<td>14</td>
<td>32</td>
<td>25.0 (0.016)</td>
</tr>
<tr>
<td></td>
<td>HS Percentile</td>
<td>64</td>
<td>42</td>
<td>100</td>
<td>86.9 (0.741)</td>
</tr>
</tbody>
</table>

2. Are there any pre-existing differences in students’ understanding of basic concepts in Newtonian mechanics prior to the fall 2003 semester?

Each student took a Force Concept Inventory (FCI) pre-test\(^c\) at the beginning of the semester. The distributions of the FCI pre-test scores are shown in Figure 4. The average scores are:

\[
TR = 16.0±0.5 \ (N = 127)^d \quad \text{vs.} \quad VP = 15.7±0.6 \ (N = 92)
\]

This shows homogeneity of two groups. Each error is calculated as \(\sigma/\sqrt{N}\), where \(\sigma\) is a nonbiased standard deviation of the corresponding distribution. For a comparison,

Honors VP = 22.7±0.2 \ (N = 35)

Average nationwide scores vary 7.3 to 20.3 (or 25% to 70%) for the pre-test. A score of 25.2 (or 87%) is considered “mastery”; 17.4 (or 60%) is the threshold for understanding Newtonian mechanics, yet fewer than 30% of physics students, nationwide meet this threshold.

\(^c\) The Force Concept Inventory (FCI) is a multiple-choice “test” designed to assess student understanding of the most basic concepts in Newtonian mechanics. The test score is a well-validated and well-accepted world wide measure of students’ conceptual understanding in mechanics. Dr. Toback implemented the FCI test (1992 version) in WebCT.

\(^d\) One student in Dr. A’s class didn’t take the FCI pre-test. Therefore \(N = 127\) instead of 128.
Figure 4: FCI pre-test score distributions for TR ($N = 127$) and VP ($N = 92$) groups, whose average scores are $16.0 \pm 0.5$ and $15.7 \pm 0.6$, respectively. Each error is calculated as $\sigma / \sqrt{N}$. $\sigma$ is a nonbiased standard deviation of the corresponding distribution.

We also evaluated how two distributions are similar in its shape. Each distribution ($f_i$) is normalized, and its error ($\delta_i$) is calculated:

$$f_i = \frac{n_i}{N} \quad \text{and} \quad \delta_i = \sqrt{\frac{(1-f_i)f_i}{N}}$$

Here $i$ is a bin number in a histogram and $n_i$ is the number of students in $i$-th bin. The similarity $S$ is defined as

$$S = \frac{\chi^2}{N_{\text{bin}}},$$

where

$$\chi^2 = \sum_i \left( \frac{(f_{TR,i} - f_{VP,i})}{\delta_i^2} \right)^2.$$ 

The results are summarized in Table 3. The values of $S$ are stable at around 1, indicating the two distributions are similar in shape within their statistical errors. Based on those data, we concluded that there is no difference in students’ academic preparation between TR0/TR1/TR2 group and VP0/VP1/VP2 group.
Table 3: Similarity of two FCI pre-test score distributions for TR and VP groups. Since the statistics are limited, we calculate $S$ for various choice of the number of histogram bins ($N_{\text{bin}}$). Each bin is defined as $[x_{\text{max}} - i \cdot \Delta x, x_{\text{max}} - (i-1) \cdot \Delta x]$, where $i = 1, 2, \cdots, N_{\text{bin}}$. $x_{\text{max}}$ is the maximum value of axis in histogram; $N_{\text{bin}}$ is the number of bins with non-zero entries.

<table>
<thead>
<tr>
<th>$x_{\text{max}}$</th>
<th>$\Delta x$</th>
<th>$N_{\text{bin}}$</th>
<th>FCI pre-test $S = \chi^2 / N_{\text{bin}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>2</td>
<td>13</td>
<td>1.09</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td>7</td>
<td>0.93</td>
</tr>
<tr>
<td>30</td>
<td>6</td>
<td>5</td>
<td>1.23</td>
</tr>
</tbody>
</table>

3. Does the VP technical writing component ask too much from students?

From the beginning, there was concern about an extensive workload in the VP program: students were encouraged to attend a one-hour TW class each week to receive instruction over writing their lab reports. One way to determine how extensive a required workload for a course is would be to compare the number of Q-drops during the semester of students in VP with those of students in TR. Given the additional work required from the students in preparing the required reports and attending the optional weekly meetings, we did not see an increase in Q-drops. Table 4 is a summary of number of Q-drop students.

Table 4: Number of Q-drop students, compared to the number of students who completed the PHYS-218 course in Table 1.

<table>
<thead>
<tr>
<th>Instructor</th>
<th>VP</th>
<th>TR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Completed</td>
<td>Q-drop</td>
</tr>
<tr>
<td>A</td>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>14 (15% of 92)</td>
</tr>
<tr>
<td></td>
<td>58 with TW</td>
<td>9 (16% of 58)</td>
</tr>
</tbody>
</table>

4. How does the VP program help the students’ final letter grades?

Our analysis addresses two questions: (i) How does the TW instruction help the students within the VP program? (ii) How does the VP lab compare to TR lab? These can be evaluated by looking at student’s final letter grades. Figure 5 shows the letter grade distributions for two groups: (a) VP1/VP2 vs. VP0 and (b) VP0/VP1/VP2 vs. TR0/TR1/TR2.

Although these statistics are limited, preliminary results show that VP with TW instruction helped to improve the students’ grades. This is also seen in a lower percentage of students who received $D$’s and $F$’s.
Figure 5: Final letter-grade distributions for students in various groups (a) VP1/VP2 vs. VP0; (b) VP0/VP1/VP2 vs. TR0/TR1/TR2.
5. Does the VP program improve exam scores?

Does the VP program enhance the students’ scores for the three mid-term exams and the final exam? The main topics in each exam include the following:

- Exam 1 - Motion in 1 and 2 dimensions;
- Exam 2 - Newton’s law, Work and Energy;
- Exam 3 - Momentum conservation, rotational kinematics/dynamics;
- Final Exam - Comprehensive coverage including simple harmonic oscillation.

Exam 1 consists of the simplest problems (i.e., kinematics) in PHYS-218, while the Final Exam consists of the most complex problems. The average score of each exam shows a trend. VP lab and recitation styles should have impacted all exams equally for VP sections, while the TW should have affected context-rich exams [6]. It should be noted that three professors’ lectures and exams are not necessarily constructed to enhance students’ conceptual understanding, since the VP program in the fall 2003 semester was a reform of recitations and laboratory experiments only.

5.1. Overall look at three professor’s classes

The average score of each exam for each group in Drs. A, B, and C’s classes is listed in Table 5. Three observations are:

(i) Dr. A’s Exam 2 and Exam 3 scores were systematically higher than other professors’ exam scores;
(ii) Drs. A and C’s VP group consistently scored higher, compared to their TR groups, while Dr. B’s VP and TR groups were virtually identical in their exam scores;
(iii) Dr. C’s VP group showed a dramatic improvement in the Final Exam, compared to his TR group.

In order to assess Observation (i), two sets of nearly identical problems for Final Exam were administrated and given to Drs. A and B’s classes. We observed a smaller difference in the average of Final Exam score between two classes. This suggests that Dr. A’s midterm exams are designed to have higher scores.

Observation (ii) can be seen in Figure 6, which is a correlation plot between Exam 1 and Final Exam scores. Most of students in Drs. A and C’s VP groups scored higher than 55% in the Final Exam, while many students in their TR groups did poorly. Both VP and TR groups in Dr. B’s class are virtually identical in their exam improvements. We speculate that this is because Dr. B’s exam style/format varied from one to another. Students tended to study the professor’s exam style and adjust their learning style accordingly. This skews the measurement of the exam improvement with time.

The improvement in Observation (iii) can possibly be explained by an impact of TW along with suitable exam style/format: Dr. C’s exams were suitable for VP (with TW) groups, while Dr. A’s exams were suitable for both VP (without TW) and TR groups. We analyze Dr. C’s exams in more detail in the following subsection.

---

Dr. A let the students know very clearly that a significant fraction of problems on the exams will be similar (to a larger or smaller extent) to some of the practice problems that he distributed before each exam. Those practice problems were the most representative problems from each chapter. The problems were not easier, but they looked familiar to the students.
Table 5: Average scores of FCI pre-test (max. 29) and exams (max. 100). Each error is calculated as $\sigma / \sqrt{N}$.

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Group</th>
<th>FCI pre-test</th>
<th>Exam 1</th>
<th>Exam 2</th>
<th>Exam 3</th>
<th>Final Exam</th>
<th>Final Exam - Exam 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>TR0</td>
<td>15.7±0.8</td>
<td>74.1±2.2</td>
<td>79.6</td>
<td>72.1</td>
<td>62.9±2.5</td>
<td>−11.2</td>
</tr>
<tr>
<td></td>
<td>VP0</td>
<td>15.0±0.9</td>
<td>79.1±2.3</td>
<td>80.5</td>
<td>78.4</td>
<td>68.8±2.2</td>
<td>−10.3</td>
</tr>
<tr>
<td>B</td>
<td>TR1</td>
<td>16.2±0.9</td>
<td>72.5±2.9</td>
<td>63.6</td>
<td>65.8</td>
<td>64.3±2.8</td>
<td>−8.2</td>
</tr>
<tr>
<td></td>
<td>VP1</td>
<td>15.3±1.0</td>
<td>71.7±3.6</td>
<td>63.8</td>
<td>67.8</td>
<td>60.2±2.7</td>
<td>−11.5</td>
</tr>
<tr>
<td>C</td>
<td>TR2</td>
<td>16.4±1.1</td>
<td>82.0±2.8</td>
<td>66.0</td>
<td>64.8</td>
<td>63.7±3.3</td>
<td>−18.3</td>
</tr>
<tr>
<td></td>
<td>VP2</td>
<td>17.1±1.3</td>
<td>82.2±3.2</td>
<td>75.1</td>
<td>70.2</td>
<td>75.5±2.3</td>
<td>−6.7</td>
</tr>
</tbody>
</table>

a) Exam style: Drs. A and C gave their exams in the most consistent style. Dr. B’s exam style was variable on each exam. However, Exam 3 and Final Exam were consistent with A and C, while Exam 1 and Exam 2 were markedly different. This difference in style skews measurement of exam improvement over time.

b) Exam context: Dr. C’s exams were suitable for VP (with TW) groups, while Dr. A’s exams were suitable for both VP (without TW) and TR groups.

Note that one student in (a), indicated by a dashed (blue) circle, has 27 (93%) for both FCI pre- and post-tests, 87, 95, and 94.5 for Exam 2, Exam 3, and Final Exam, respectively. Therefore, the student’s Exam 1 score does not reflect the student’s academic performance in PHYS 218.

Figure 6: Exam 1 vs. Final Exam for (a) Dr. A’s class, (b) Dr. B’s class and (c) Dr. C’s class.
5.2. Dr. C’s class

As mentioned earlier, Dr. C’s exams were consistent in style, making them suitable for measuring the impact of TW. It should be noted that, despite the extra requirement of TW sessions and writing lab reports, the VP2 students virtually carried out the HW assignments at the same level as the TR students as shown in Table 6.

Figure 7 shows the change in the average score of each exam (Exam 2, 3, and Final Exam) relative to Exam 1 for his VP2 and TR2 groups. The average scores of Exam 1 from VP and TR groups were 82.2 and 82.0, indicating the initial academic backgrounds are approximately equal between the two groups. This is consistent with the analysis in the previous section. We saw a dramatic improvement, however, for VP compared to TR by the end of semester. Is this improvement due to a level of completing HW assignments? Is this improvement due to changes in his lecture during 2003?

Table 6: Number of students who completed HW assignments through WebCT.

<table>
<thead>
<tr>
<th></th>
<th>100%</th>
<th>80%-99.9%</th>
<th>60%-79.9%</th>
<th>&lt;60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP2 (N = 26)</td>
<td>13</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>TR2 (N = 30)</td>
<td>14</td>
<td>3</td>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>

Figure 7: Change of the average score of each exam for students in VP2 (Visual Physics Group) and TR2 (Control Group) in Dr. C’s PHYS-218 class in fall 2003 semester, relative to Exam 1. \( X = 1, 2, 3, \) and 4 correspond to Exam 1, 2, 3 and Final Exam. Note that the average scores of Exam 1 from VP2 and TR2 groups are 82.2 and 82.0. Each error bar is calculated as \( \sigma / \sqrt{N} \).
Dr. C has taught PHYS-218 class for three fall semesters in 2001, 2002, and 2003 in a consistent style: using the same textbook, using PowerPoint for his lectures, and using the same style for exams. Table 7 shows a comparison of the average score of each exam. Exam 1 is always the highest score. This is because it tested the simplest of physics concepts using a minimum level of math skills. The Final Exam score is consistently lower because it tests all physics concepts comprehensively.

Table 7: Average scores in Dr. C’s exams.

<table>
<thead>
<tr>
<th>Semester</th>
<th>Group</th>
<th>Exam 1</th>
<th>Exam 2</th>
<th>Exam 3</th>
<th>Final Exam</th>
<th>Difference from Exam 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2001</td>
<td>TR (N = 105)</td>
<td>78.9</td>
<td>63.6</td>
<td>64.9</td>
<td>60.1</td>
<td>−18.8</td>
</tr>
<tr>
<td>Fall 2002</td>
<td>TR (N = 93)</td>
<td>84.7</td>
<td>75.5</td>
<td>69.7</td>
<td>73.6</td>
<td>−11.1</td>
</tr>
<tr>
<td>Fall 2003</td>
<td>TR (N = 30)</td>
<td>82.0±2.8</td>
<td>66.0</td>
<td>64.8</td>
<td>63.7±3.3</td>
<td>−18.3</td>
</tr>
<tr>
<td></td>
<td>VP (N = 26)</td>
<td>82.2±3.2</td>
<td>75.1</td>
<td>70.2</td>
<td>75.5±2.3</td>
<td>−6.7</td>
</tr>
</tbody>
</table>

We should expect that Exam 1 scores reflect the least impact of VP and TR labs, where the Final Exam scores reflect the most impact. Therefore, effectiveness in teaching physics can be assessed from a correlation between those exam scores. The distributions of Exam 1 scores are shown in Figure 8(a) and Figure 8(b). The shapes of the distributions are similar, again indicating the initial academic backgrounds are approximately equal between the two groups. However, as in Figure 9(a) and Figure 9(b), the shapes of the distributions in Final Exam are different. The data show that very few students in VP scored low (e.g., below 60) on the Final Exam. Table 8 is similar to Table 3 but is based on Exam 1 (Figure 8) and Final Exam (Figure 9) distributions. If two distributions are similar within its statistical fluctuation, we expect to see the values of S to be around 1. It is clear that two distributions in Final Exam are different, especially at 60%. Thus, the improvement in Final Exam can be illustrated as in Figure 10.

Figure 8: [Fall 2003 Semester] Distributions of Exam 1 scores for TR2 (control) and VP2 groups in Dr. C’s class are shown in (a) and (b), respectively. The shapes of the distributions are similar, indicating the initial academic backgrounds are approximately equal between the two groups.
Figure 9: [Fall 2003 Semester] Distributions of Final Exam scores for TR2 (control) and VP2 groups in Dr. C’s class are shown in (a) and (b), whose average scores are 63.7(±3.3) and 75.5(±2.3), respectively. The error in each parenthesis is calculated as $\sigma/\sqrt{N}$. Please notice the difference in a fraction of students whose Final Exam scores are higher than 60%.

Table 8: Similarity of two exam distributions for TR and VP groups. Since the statistics are limited, we calculate $S$ for various choice of the number of histogram bins ($N_{\text{bin}}$). Each bin is defined as $[x_{\text{max}} - i \cdot \Delta x, x_{\text{max}} - (i-1) \cdot \Delta x]$, where $i = 1, 2 \ldots, N_{\text{bin}}$; $x_{\text{max}}$ is the maximum value of axis in histogram; $N_{\text{bin}}$ is the number of bins with non-zero entries.

<table>
<thead>
<tr>
<th>$x_{\text{max}}$</th>
<th>$\Delta x$</th>
<th>$N_{\text{bin}}$</th>
<th>$S = \chi^2/N_{\text{bin}}$</th>
<th>$x_{\text{max}}$</th>
<th>$\Delta x$</th>
<th>$N_{\text{bin}}$</th>
<th>$S = \chi^2/N_{\text{bin}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>6</td>
<td>0.69</td>
<td>100</td>
<td>10</td>
<td>8</td>
<td>1.4</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>3</td>
<td>0.85</td>
<td>100</td>
<td>20</td>
<td>4</td>
<td>2.3</td>
</tr>
<tr>
<td>100 &gt;80, &lt;80</td>
<td>2</td>
<td></td>
<td>0.89</td>
<td>100 &gt;60, &lt;60</td>
<td>2</td>
<td></td>
<td>7.3</td>
</tr>
</tbody>
</table>
Figure 10: Illustration of improvements of Final Exam scores in the VP (with TW instruction) program, compared to our traditional approach.
In the next stage, we analyzed a correlation plot between Exam 1 scores and Final Exam scores in each of three fall semesters in 2001, 2002 and 2003. See Figure 11 below. No significant change is seen for students who received the same TR lab for all three semesters. Further, one can examine a fraction of students whose Final exam scores are higher than 60%. Is the correlation seen in VP2 in Figure 11(c) (same as in Figure 6(c)) a real effect or just a fluctuation?

Figure 11: Correlation between Exam 1 and Final Exam scores in Dr. C’s PHYS-218 classes for three fall semesters in 2001, 2002 and 2003. Students in VP lab with TW session and TR lab are shown in green solid circles and red solid triangles. Note a fraction of students whose Final Exam scores are higher than 60%.

A way to test whether the correlation is due to a statistical fluctuation is to re-plot Figure 11(a) and Figure 11(b), but with groups with smaller number of students. We chose the first 2 sections as one group and the last 2 sections as another in each semester. Figure 12(a) and Figure 12(b) are the correlation plots for the fall 2001 semester; Figure 12(c) and Figure 12(d) are for the fall 2002 semester. None of four plots look closer to the one seen in VP2.

In summary, we observe a significant impact of the VP (with TW instruction) program at least for Dr. C’s class in the fall 2003 semester.
Figure 12: [Fall 2001 and 2002 Semesters] Same as in Figure 11, but showing each group of two sections, simulating a similar size of students in VP2 group in the fall 2003 semester. All students in (a)-(d) took the TR labs. Notice that in each plot there is not-small fraction of students whose Final Exam scores are lower than 60%.

6. What is the impact of TW instruction on writing the scientific lab report?

One of goals in the VP program is to teach students how to write a scientific lab report in the early stages of their academic careers. The Center for Teaching Excellence independently evaluated two sets of lab reports from three groups:

Two lab reports: Pre-test reports were submitted after the students received 2-3 sessions of TW instruction; therefore, even for the pre-test, the students had received some TW instruction. The post-test reports were submitted when TW instruction had concluded.
Three groups: For non-honors, Dr. A’s VP group (VP0) was not required to attend TW class, while Drs. B and C’s VP groups (VP1 and VP2) were required. In addition, honors with TW were added as the 3rd group (VPHonor) for this analysis.

For an assessment such as this, it is key to prepare two VP groups of non-honors students whose academic backgrounds are identical. Measurement of Research Services measured their academic preparation before taking PHYS 218 course by looking at their HS class rankings and their SATV scores. The results show that there were no pre-existing differences among the three groups on HS ranking. The honors group was significantly higher than the others in SATV score. The students in the VP0 group had slightly higher SATV scores than those in the VP1/VP2 group, but not significantly.

Figure 13 shows the results of evaluation of the lab reports based analytically on a 100-point scale. It is natural that students in the honors sections scored higher than students in VP0, VP1, and VP2 groups. However, their TW skills, as reflected by their scores, did improve dramatically. Further, the group not receiving TW instruction, VP0 group, showed no improvement between pre- and post-lab reports.

SATV scores and its significance in the difference of mean scores for three groups are summarized in Table 9, along with their Pre- and Post-test scores. Considering the HS ranking and SATV scores of students in the VP1/VP2 group, improvement in the quality of scientific lab reports is dramatic even after they received 2-3 sessions of TW instruction (i.e., as is reflected in the pre-test score).
Table 9: SATV, pre- and post-test scores for each of three groups. The significance between VP0 (or VPHonors) and VP1/VP2 in Scheffe Post Hoc tests is shown.

<table>
<thead>
<tr>
<th></th>
<th>SATV</th>
<th>Significance</th>
<th>Pre-test</th>
<th>Significance</th>
<th>Post-test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP0</td>
<td>623.6</td>
<td>0.130</td>
<td>42.88</td>
<td>0.004</td>
<td>47.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VP1/VP2</td>
<td>591.1</td>
<td>&lt;0.001</td>
<td>53.12</td>
<td>0.045</td>
<td>68.58</td>
<td>0.184</td>
</tr>
<tr>
<td>VPHonors</td>
<td>679.8</td>
<td></td>
<td>60.51</td>
<td></td>
<td>75.24</td>
<td></td>
</tr>
</tbody>
</table>

With further correlation analysis of the change between pre-test and post-test scores, it is found that:

(i) HS ranking and SATV scores were not significantly correlated with the changes. Together they accounted for about 2% of the variance;

(ii) Considering students in the VP1/VP2 and VPHonors groups, the gain in writing is correlated with the number of hours of instruction. Pearson product-moment correlation is 0.38. This correlation would likely be higher except that, within the treatment, there was very little variability in hours of instruction. The mean number of hours of instruction was 11.3 and the median was 12 out of a maximum of 13 hours. Seventy-five percent of students in the groups had 11 or more hours of instruction.

The preliminary results of evaluation of the lab reports based holistically on a 4-point scale are also shown in Figure 14. A 4-point report is considered as excellent, while a 1-point report as poor. In order to evaluate the numerical improvement in the holistically analyzed data, a fraction of number of reports with excellent (4 points) and good (3 points) assessments, $f$, is calculated:

$$f = \frac{N_{excellent} + N_{good}}{N_{excellent} + N_{good} + N_{fair} + N_{poor}}$$

The group not receiving TW instruction, VP0 group, showed little change in $f (3.0\% \rightarrow 6.1\%)$ between pre- and post-lab reports, while a dramatic increase in producing excellent and good reports is seen with TW instruction. Both analyses clearly show the importance of the TW instruction (as defined by lectures, class participation, and feedback on the reports). It is important to note that simply requiring the students to write lab reports while not receiving TW feedback or instruction, as was the case with the VP0 group, was not enough to affect post-test performance.
7. Were the interactive-engagement methods effective? If so, how?

The superior efficacy of cooperative learning approaches has been documented in a variety of studies over several decades of physics education research. Hake [7], in a study of almost 6000 students in physics mechanics courses, shows that the use of interactive engagement (IE) methods, as part of a holistic and articulated physics course reform effort results in higher conceptual gains, as measured by pre- and post-FCI scores than traditional lectures.

IE is achieved by Socratic-like questioning of students with the TA acting as “coach,” challenging students to face complexity and ambiguity that requires a cooperation among students as they work in groups of three and apply logical chain of reasoning to solve context rich problems [7]. Students interact with each other, with the instructor as a coach or guide, and with materials synthesized from several key physics principles and concepts, folding into one complex scenario [2]. A key ingredient is frequent and thoughtful interaction as well as Just-in-Time intervention, by the TA [8] when students are stuck. Most
of the quantitative research in science education regarding the effectiveness of IE in a learning environment comes from a rich history of models and methods generated by the physics reformed teaching community. Hake [9, 7] defines IE methods as those designed at least in part to promote conceptual understanding through interactive engagement of physics students in heads-on (always) and hands-on (usually) problem solving, which yield immediate feedback through interaction with peers and instructors.

The physics TAs were instructed in a cognitive apprenticeship model of instruction and applied methods of “cognitive coaching to the recitation and lab” [10]. The Reformed Teaching Protocol Assessment of TAs [11] was periodically given during recitation and/or lab by the physics education researcher, C. Ezrailson. Scoring each of these 25 items from 0 - 4 results in a lesson score ranging from 0 - 100 points describing the degree of reform present in a lesson [12]. Figure 15 shows the scores as a function of assessment numbers. For physics lessons we have observed, some typical approximate scores are:

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional university lecture (passive)</td>
<td>&lt; 20</td>
</tr>
<tr>
<td>University lecture with demonstrations (some student participation)</td>
<td>&lt; 30</td>
</tr>
<tr>
<td>Traditional high school physics lecture (with student questions)</td>
<td>&lt; 45</td>
</tr>
<tr>
<td>Partial HS reform (some group-work; most discourse still with teacher)</td>
<td>&lt; 55</td>
</tr>
<tr>
<td>Medium sized (N &gt; 50) university lectures with Mazur-like group-work (ConcepTests)</td>
<td>55-75</td>
</tr>
<tr>
<td>and a student Personal Response System</td>
<td></td>
</tr>
<tr>
<td>Modeling curriculum (varies with amount and quality of discourse)</td>
<td>75-99</td>
</tr>
</tbody>
</table>

Figure 15: Mean scores of Reformed Teaching Observation Protocol (RTOP) evaluation on interactive teaching for TAs in VP group and TAs in TR (control) group [12]. The evaluation (= assessment) is performed approximately every two weeks. The RTOP score 50 for VP TAs at assessment number 0
(zero) was obtained before the training was given (one week before the first week of the semester). Note that the score is expected to be similar to the score (= 44) for TR TAs at assessment number 1.

In addition, students were given the recitation survey instrument (evaluation surveys) periodically during the semester in order to assess the interactive engagement methods used by their TAs and acted as interim measures of success during recitation. Our student survey instrument is based upon the RTOP instrument [11] with AAAS, NCTM, NRC standards; constructivist, collaborative inquiry and active teaching methodology. The RTOP establishes an operational definition for reformed physics teaching by quantifying twenty-five observable behaviors each of which is scored from 0 - 4 as follows:

0. the behavior never occurred
1. the behavior occurred at least once
2. occurred more than once; very loosely describes the lesson
3. a frequent behavior, fairly descriptive of the lesson
4. pervasive or extremely descriptive of the lesson

Figure 16 shows the mean of median survey scores for TR and VP group in Weeks 1, 7 and 14 of PHYS-218 course. The VP students’ scores show students’ satisfaction with the methods used and reflect how well the recitation and lab adhered to the interactive engagement method used by their TAs. It should be noted that, as we expected, the average scores in Week 1 are nearly identical, indicating there was no difference in students’ observation on recitation and lab method at the beginning of the semester. The difference between TR and VP students’ scores increased as the semester progressed.

**Figure 16:** Mean value of median scores of student recitation surveys (analyzed in 0-4 point scale) in week 1, 7 and 14. Higher value reflects high student satisfaction and confidence in the interactive-
engagement methods as promoting conceptual understanding of physics in both recitation and lab. The student surveys correlate with the RTOP assessment items, showing agreement between the TA evaluation and the students’ assessment of TA adherence to reformed methodology.

8. Does the VP program improve students’ conceptual understanding?

Each student was asked to take FCI tests (“pre-test” and “post-test”) at the beginning and end of the semester. The FCI score is a well-validated and well-accepted worldwide measure of a student’s conceptual understanding in mechanics, not computational knowledge or factual knowledge of physics.

During the fall 2003 semester, the reform was limited to the recitation (30 minutes of 1 hour period) and laboratory experiments. The lectures by three instructors were more or less traditionally given. Therefore, it is expected that an impact on students’ conceptual understanding by the VP program should be equal across students in VP0/VP1/VP2 groups, and that improvement was modest as conceptual understanding of physics was tenacious to change and occurs with much repetition over time. During recitation student understanding was promoted through the solution of complex concept, context-rich problems. However, it is interesting to analyze the FCI data to see students’ beginnings in the development of the conceptual understanding of physics concepts.

Not all of students took both tests: 108 out of 129 students for TR group and 77 out of 92 for VP group. We calculated the average FCI scores for students who took both pre- and post-tests, and summarize them in Table 10. A score of 25.2 (or 87%) is considered “mastery”; 17.4 (or 60%) is the threshold for understanding Newtonian mechanics, (although fewer than 30% of introductory physics students, nationwide are able to reach even the threshold level). The majority of our VP students were at or near the threshold at the end of semester. Pearson correlation between FCI pre- and post- test scores (see Figure 17) is 0.660 for the VP group; 0.633 for the TR group. A value of 0.5 or higher indicates a strong correlation expected from these measurements.

Table 10: Average scores (max. = 29) for the Force Concept Inventory (FCI) tests at the beginning and end of the fall 2003 semester. Each error is calculated as $\sigma/\sqrt{N}$. $\Delta$ is the difference of FCI scores between pre-test and post-test.

<table>
<thead>
<tr>
<th>Instructor</th>
<th>For Students who took both pre- and post-tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FCI pre-test (N)</td>
</tr>
<tr>
<td>A</td>
<td>15.7±0.8 (51)</td>
</tr>
<tr>
<td></td>
<td>15.0±0.9 (34)</td>
</tr>
<tr>
<td>B</td>
<td>16.2±0.9 (46)</td>
</tr>
<tr>
<td></td>
<td>15.3±1.1 (32)</td>
</tr>
<tr>
<td>C</td>
<td>16.4±1.1 (30)</td>
</tr>
<tr>
<td></td>
<td>17.1±1.3 (26)</td>
</tr>
<tr>
<td>All</td>
<td>16.0±0.5 (127)</td>
</tr>
<tr>
<td></td>
<td>15.7±0.6 (92)</td>
</tr>
</tbody>
</table>
Figure 17: FCI post-test scores vs. pre-test scores. Pearson correlation between FCI pre- and post- test scores is 0.633 for the TR group ($N = 108$) and 0.660 for the VP group ($N = 77$).

We noticed that a large fraction of students in Dr. C’s class did not take the FCI post-test. Since the test was given to students who completed HW assignments, a question was raised: was the dramatic improvement in the Final Exam due to the VP2 students’ spending their time preparing for the Final Exam instead of HW assignments? Table 11 is a summary of the FCI test scores and Final Exam scores for students who completed HW assignments and for those who didn’t. It is clear that the students who completed both HW assignments and VP program achieved higher scores than any other student groups.

Table 11: [Dr. C’s students] Average scores (max. = 29) and its $\sigma$ values for the Force Concept Inventory (FCI) tests at the beginning and end of the fall 2003 semester, comparing to the Final Exam scores.

<table>
<thead>
<tr>
<th>FCI post-test, completed?</th>
<th>Group</th>
<th>FCI pre-test</th>
<th>FCI post-test</th>
<th>Final Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>TR2 ($N = 17$)</td>
<td>15.2±5.4</td>
<td>18.3±4.9</td>
<td>69.2±12.8</td>
</tr>
<tr>
<td></td>
<td>VP2 ($N = 17$)</td>
<td>16.6±6.8</td>
<td>20.1±4.4</td>
<td>78.6± 9.9</td>
</tr>
<tr>
<td>NO</td>
<td>TR2 ($N = 13$)</td>
<td>17.9±6.8</td>
<td>n/a</td>
<td>56.5±21.4</td>
</tr>
<tr>
<td></td>
<td>VP2 ($N = 9$)</td>
<td>18.0±6.7</td>
<td>n/a</td>
<td>69.6±11.5</td>
</tr>
</tbody>
</table>

8.1. Hake gain

For a larger group, Hake gain, $g$ (a fraction of the maximum possible gain realized [13]), is one of quantities to measure an effectiveness of interactive engagement in this program. It is defined as

$$g = \frac{\Delta}{29 - \langle p_{\text{pre}} \rangle},$$

where

$$\Delta = \frac{\sum_i (p_{\text{post},i} - p_{\text{pre},i})}{N}$$

and

$$\langle p_{\text{pre}} \rangle = \frac{\sum_i p_{\text{pre},i}}{N}.$$ 

Here $p_{\text{post},i} - p_{\text{pre},i}$ is the change of the pre-test score to the post-test score for student $i$. 
Table 10 also shows Hake gains for VP and TR groups. Our data shows a small gain \((g \sim 0.2)\) in VP group and negligible effect \((g \sim 0.1)\), reflecting traditional methods for TR group. Our data points are placed in the Hake gain \(g\) vs. FCI pre-test score (%) plane in Figure 18, taken from Ref. [7], where \(g > 0.3\) is expected for classes that are taught in more interactive manner [12]. The VP point is a lower edge of a band expected in the other research [7].

It should be noted that the VP program is a reform of recitation and laboratory experiment with TW instruction in order to improve students’ thinking about physics and grasp of physics concepts. In the fall 2003 semester, however, one-hour recitation was spent splitting 50-50 between traditional homework problems from the textbook and context-rich problems specially prepared for VP program. In the lecture, students were taught traditionally. It is important to note that a systematic reform, encompassing more than 30 minutes of application of these methods of context-rich problems each week alone, is required in order to more dramatically improve students’ conceptual understanding. The gain \((g = 18\%)\) in the VP group, however, compared to the gain \((g = 9\%)\) in the TR group, was significant as seen in Figure 18. A greater change between pre- and post-test scores on FCI will be expected with a further reform that also encompasses, lecture, homework and exams by professors as well as an extension of interactive engagement recitation techniques to interactive labs, web support, etc. Those were not a scope in the fall 2003 semester.
**Figure 18:** Hake gain $g$ vs. FCI pre-test score (%) taken from Ref. [7]. Our VP and TR data points are shown in + (green plus) and X (red cross), respectively.

### 8.2. Cohen’s $d$

Cohen's $d$ is a standardized measure and commonly used to calculate a size of the effect. It is defined:

$$d = \frac{\langle p_{\text{post}} \rangle - \langle p_{\text{pre}} \rangle}{\sigma_{\text{pooled}}}$$

where $\sigma_{\text{pooled}} = \sqrt{(\sigma_{\text{post}}^2 + \sigma_{\text{pre}}^2)/2}$. Here $\langle p_{\text{pre}} \rangle$ and $\langle p_{\text{post}} \rangle$ are the mean values for two measurements (e.g., FCI pre- and post-test) for a group; $\sigma_{\text{pre}}$ and $\sigma_{\text{post}}$ are its standard deviations; $\sigma_{\text{pooled}}$ is the standard deviation adjusted for difference in the sample size and known as the "pooled" estimate of the standard
deviation. See Figure 19. An effect size of 0.25 or more is commonly considered to be practically and educationally significant [14].

Figure 19: Cohen’s $d$ in measuring an effect size of the change between FCI pre- and post-test scores for each of VP group ($N = 77$) and TR group ($N = 108$). An effect size of 0.25 or bigger is considered to be statistically significant.
IV. Conclusion: Summary of Results

In the fall 2003 semester, we introduced the VP program, a reform of recitation and laboratory experiment with TW instruction, to non-honors students in PHYS-218 course. The goal of the program was to provide a coherent effort to improve students’ understanding of physics (i.e., exams), writing ability of scientific papers (i.e., lab reports), and a teamwork in the early stages of their academic careers. Although the statistics are limited, the results were positive because we made an extra effort in improving teaching. The data shows that students’ learning was enhanced with TW instruction.

However, it was small improvement in student’s mastery of conceptual understanding based on FCI test. This suggests that a reform of our lecture, as a part of a total reform effort, is necessary to be more consistent with context-rich exams, since recitation was expected to do so effectively in 30 min during the one-hour weekly recitation time.

A summary of the most significant results of our research is provided in the answers to our guiding questions:

1. Are there any pre-existing differences in students’ academic backgrounds prior to the fall 2003 semester that would affect their performances in VP?
   No. We conclude that homogeneity of three groups was acceptable for our assessments.

2. Are there any pre-existing differences in students’ understanding of basic concepts in Newtonian mechanics prior to the fall 2003 semester?
   No. We conclude that the homogeneity of three groups was acceptable for our assessments.

3. Does the VP technical writing component ask too much from students?
   No. However, the technical writing required extensive time for many students. We plan to re-organize the technical writing component so that the amount of time needed to complete the assignments is reduced.

4. How does the VP program help the students’ final letter grades?
   This program reduced the number of D’s and F’s students receive. Further, VP with TW instruction helped to increase the number of B’s and A’s students receive.

5. Does the VP program improve exam scores?
   Although the statistics are limited, the VP lab, especially with rigorous writing component, improved the students’ exam scores toward the Final Exam. We note that one class had no effect. We speculated that a frequent change in the exam style/format confused students in the class as they tended to study the professor’s exam style and adjust their learning style accordingly.

6. What is the impact of TW instruction on writing the scientific lab report?
   The gain in writing is correlated with the number of hours of instruction. The analyses clearly show the importance of the TW instruction (as defined by lectures, class participation, and feedback on the reports). It is important to note that simply requiring the students to write lab reports while not receiving TW feedback or instruction, as was the case with the VP0 group, was not enough to affect post-test performance.
7. Were the interactive-engagement methods effective? If so, how?
   Yes. The success of interactive engagement methods, modeled in training and adhered to by VP TAs during recitation problem solving, was reflected in higher conceptual gain, as measured by pre- and post-FCI scores, even if the lectures were taught traditionally.

8. Does the VP program improve students’ conceptual understanding?
   Yes. The gain ($g = 18\%$) in the VP group, compared to the gain ($g = 9\%$) in the TR group, was significant as seen in Figure 18. It should be noted that in the fall 2003 semester, one-hour recitation was spent splitting 50-50 between traditional homework problems from the textbook and context-rich problems specially prepared for VP program. A greater change between pre- and post-test scores on FCI will be expected with a further reform that encompasses lecture, homework and exams by professors as well as an extension of interactive engagement recitation techniques to interactive labs, web support, etc. Those were not a scope in the fall 2003 semester.

   We also plan to monitor VP students from fall 2003 in PHYS-208 in spring 2004. We are looking forward to seeing more comprehensive results in the future semesters.

Acknowledgements

   We thank the Department of Physics for financial support of the visual physics program, and Drs. Lewis Ford and Ed Fry for their encouragement.
Appendix A: Visual Physics – the details

Recitation. Students are grouped into teams of three for both recitation and lab, and both activities are conducted in the lab room. In recitation, VP uses the tools of cooperative learning groups and context-rich problems to build student conceptual learning and problem-solving skills. The TA begins the recitation session by asking if any students had difficulties with particular homework problems on the current assignment. From the responses, the TA spends some time explaining the problem-solving strategy and mode of thinking needed to solve the problems in question. Then the TA assigns a context-rich problem [6] and the students are given approximately 20 minutes to solve the problem. Each student team works as a team on the problem. The TA circulates throughout the room, observing each group as they work. If a group is stuck in their effort, the TA intercedes at that point to question and answer the question, so that they get going again. At the end of the allotted time, the TA collects their work and repeats the process with another context-rich problem.

There are three points of departure between this strategy our traditional expository technique.

- The TA does not work problems for the students.
- The TA delineates the problem-solving strategy and the several concepts addressed in assigned homework problems with which students may have difficulty.
- During the context-rich-problem quiz, the TA intercedes as a coach to help facilitate the process if and when any team gets stuck. Pedagogic research shows that the exact moment when real learning happens most effectively occurs when students are “ready” to hear the concept delineated. The students work as teams on the quiz problems. Within the dynamics of each team, students engage in negotiation and application of what has been learned in lecture and homework review – a reinforcing learning strategy. The sticking points encountered by different teams are frequently on different specifics; by intervening at the team level the TA can provide focused instruction when those students are ready for it, then move to the next group.

Laboratory. The physics topics and apparatus of the laboratory experiments are largely the same as those in the traditional course, but the experience is wholly different. The students learn to work as a team. Each experiment is structured within three or more problems, using the same equipment to study several different questions. The work of each team is structured, with roles of doer (sets up the equipment and conducting the experiment), scribe (operates the data acquisition and logs the data), and manager (maintains oversight vis a vis objectives and makes preliminary analysis to see that results make sense). The team rotates roles on successive problems.

(i) Each problem is inquiry-driven, with opportunities and credit for minimizing and quantifying sources of uncertainty and investigating departures from expectations.

(ii) The experiments use video to enhance the visualization of kinematics and dynamics. Each setup is equipped with a space frame and a high-resolution video camera which is linked via firewire to a PC. The image sequence from the camera is acquired directly into LabView™, the most widely used data acquisition software used in both research and industry. LabView™ has a powerful intuitive toolkit, that enables the students to cursor-select points within each frame and automatically transfer the x-y screen coordinates to three columns (frame #, x, y) of an EXCEL® spreadsheet. The students construct their
analysis of the data in the spreadsheet. That enables them to easily produce graphics of their data, to perform regression fits to data, etc.

(iii) *The students develop proficiency in REAL software and hardware, not dead-end educational packages.* The software used in the course is LabView™, EXCEL®, WORD®, and AutoCad™. These are professional tools that the students will use throughout their college education and then in their careers as either scientists or engineers. We are equipping them in their first semester!

(iv) *Each student writes a serious 2-page scientific paper for each lab report.* At the end of each 2-week experiment, each student of a team is assigned a different problem to write up as a paper. The paper is written in the form a short technical paper, using the full set of format and style rules that apply to journal articles and papers for conference proceedings. The papers are submitted electronically. You can read samples of their work for yourself at http://www-english.tamu.edu/pers/fac/raisor/presentationhome.html.

(v) *The students receive instruction in and credit for technical writing.* We employ Prof. Cindy Raisor, an experienced teacher of English composition and technical writing, to teach evening classes (provided for all students!) in the many elements of technical writing pertinent to the genre of a short technical paper – the ubiquitous mode of communication among scientists and engineers. Her lectures span the gamut of what it takes to write well, from word usage and sentence structure to paragraph organization to the organization of a paper and effective use of graphics and tables. If you peruse the examples of the students’ papers at her web site, you will get the flavor of how this instruction is building the students into able writers during their first semester at TAMU. Their lab reports are graded twice – by the TA for physics content and by Prof. Raisor and her TA for technical writing.

**Technical Writing.** Our students are becoming less proficient at writing well. An increasing fraction of students are escaping the traditional core requirements of English composition and technical writing. Yet all but a few of our students write so poorly it brings tears to the eyes to read their efforts.

Heller addressed this problem at Minnesota by teaming with the English department to offer technical writing instruction in a co-registered lecture with Freshman Physics in which the writing assignments were the actual lab reports for the experiments in the physics course. Both English and Physics faculty were enthusiastic about the outcome: students were trained in writing few-page technical papers, the genre that is ubiquitous as the primary mode of communication in all fields of science and engineering. And they received that training in their freshman year, when it could benefit through the rest of their college courses. (At TAMU today, registration is so jammed for the technical writing course that typically only seniors can enroll.)
References

1. D. Toback, Computerized Homework Assignment Grading System (CHAGS). See the following URL: [http://faculty.physics.tamu.edu/toback/218/WebCT/](http://faculty.physics.tamu.edu/toback/218/WebCT/)

2. K. Heller and M. Hollabaugh, “Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups,” American Journal of Physics 60(7), 637 (1992); See, for example, the following URL: [http://groups.physics.umn.edu/physed/](http://groups.physics.umn.edu/physed/)


4. Texas A&M University Writing Center, “Goals for Writing-Intensive Courses,” retrieved from URL: [http://uwc.tamu.edu/faculty/](http://uwc.tamu.edu/faculty/)

5. [http://web.uvic.ca/psyc/coursematerial/psyc300b.s02/300B/Lectures/10_Post-Hoc_Comparisons.PDF](http://web.uvic.ca/psyc/coursematerial/psyc300b.s02/300B/Lectures/10_Post-Hoc_Comparisons.PDF)


