1. The course

We have been teaching a class in introductory statistics course at Penn State for many years. At the main (University Park) campus, the course is taught in the fall, spring, and summer semesters with an annual enrollment of about 2,200. It is a pre-calculus, introductory survey course. Until two years ago, this course was taught in a 'traditional' fashion. With the support of a grant from the Pew Foundation's Center for Academic Transformation, it was totally restructured. In the previous (traditional) format, students attended three lectures and two recitation meetings per week. Experienced, full-time faculty lectured to groups of about 240 students. Twelve graduate teaching assistants (GTAs) each taught two, one-hour recitation sections twice a week to about 40 students. GTAs also held office hours and graded exams. The traditional structure was labor-intensive, requiring many hours of faculty and GTA time per semester, creating resource problems for the department. More importantly, the traditional structure was not as effective academically as it should have been.

The redesigned course has one large group meeting (LGM) per week. The traditional recitation sections were changed to computer-mediated workshops, with technology-based independent learning materials and computerized testing added to give students more practice time and feedback, and instructional roles were shifted from information presentation to learning facilitation. The combination of smaller classes and computer-mediated data workshops enables faculty to have more one-to-one contact with individual students. Faculty are able to address the different needs of individuals, and students can be challenged according to their own skill levels. Computer-based classes enable students to work in teams, which generates more active participation. There is frequent hands-on experience with statistical analysis and with the visualization of concepts. GTA roles shifted under the new structure from instruction to guidance. Technology-based instruction and collaborative activities enable students to explore material that was formerly taught by teaching assistants. Fewer lectures allow faculty to guide most of the section meetings that were formerly led by teaching assistants. GTAs are paired with faculty and an undergraduate intern in the labs, enabling the faculty member to model ways to facilitate learning.

The most important learning goals for the course require students to:

- actively engage with course materials and other students.
- actively participate in data analysis and design;
• understand the reasoning by which findings from sample data can be extended to larger, more general populations
• understand and apply basic concepts of statistics (e.g., variables, inferences, probability, types of designs, etc.);
• design, conduct, and analyze a scientific research study, including using computer software.
• critically evaluate the results of scientific studies

2. A typical class.

There are two types of classes: Computer Labs and LGM’s (Large Group Meetings). There are two computer labs and one LGM each week, sequenced as Lab>LGM>Lab or as LGM>Lab>Lab, depending on the time the class is offered.

Computer Labs. In the first lab the instructor takes about 4 or 5 minutes to look at what will be doing during the week. This may be followed by a brief overview of the main concept covered in the lab— anywhere from five to 10 minutes. We have an excellent course management system which provides students with all of the activities we will be doing during the week, including reading and homework assignments, datasets to be used, lab quizzes scheduled, etc. (see [5] for the URL). Here is the 'This Week’ page on the web site for Week 11: March 18-22, 2002:

[This Week

Week 11: Monday, March 18, 2002 - Friday, March 22, 2002

Lab 19 March 18, 2002.

1. What To Read Sections 13.1-13.2 before today's lab. Read Sections 12.5-12.7 (covered on RAQ 3) before Wednesday's LGM and 13.3-13.5 before Friday.

2. Exercises Homework assignment 9: Chapter 12--12.17, 12.21, 12.28, 12.30-12.32, 12.35.

3. Today's lab activities will be concerned with inference for one population mean, including paired comparisons: confidence intervals and testing.

4. Datasets Use the data from the Spring Survey

5. Study Guides RAQ 3 is scheduled for Wednesday's LGM. Take a look at the Study Guide for Chapters 9, 10, 12 and 13

6. Test Pilot Take the Lab Quiz (see [6] for information about Test Pilot)

LGM Wednesday March 20, 2002
1. RAQ 3 will be given today. Please bring a #2 pencil. Coverage for the RAQ is Sections 9.3-9.7, Chapter 12, and Sections 13.1-1.2

**Lab 20, Friday March 22**

1. We need to form groups for Project II. Please login on to coursetalk (an asynchronous web-based communication software, like a chat room) and check your membership. If you are not listed, contact Yudan or Eliza (GTAs) to correct the situation. If you want to change groups, we can consider that as well. The URL is coursetalk.cac.psu.edu

2. **Activities** Today's activities will be on two proportions and two means. We will look at whether or not the proportion of students who would have sex without being in a committed relationship is the same or not for whose who have driven under the influence and those who have not. In comparing two means we will get confidence intervals and tests.

3. **Datasets** Use the data from the Spring Survey

4. Use the \( t^* \)-table of multipliers

5. One of the questions on the Lab Quiz asks you what type of general 'theme' is most appealing to you. Confer with members of your group to express your preference.

6. **Test Pilot** Take the Lab Quiz]

The schedule for Monday's lab is fairly typical. After the brief look at the week's agenda, students begin working. As soon as they enter the lab, they login on their PC, get on the Internet, bring up this week's agenda, then open Minitab (a statistical software package), go back to their 'This Week' page, move their cursor to 'Datasets', and click on 'Spring Survey', copy and paste it into their Minitab Worksheet. 'Spring Survey' refers to data collected into a data file from the students during the first week of classes. It consists of about 40 items on a variety of variables, like their gender, grade point average, race, eye color, height, weight, ideal height and weight, drinking habits, etc. This data is then used throughout the semester to illustrate statistical concepts. A student intern hired by the department passes out an 'activity' for the students to work on or it is on-line (on the web), while a teaching assistant at the front of the lab does the same things as students are to do, but on a delayed basis. The instructor walks around the room prepared to help any students having problems or questions. After the students have worked (individually, in pairs or in groups of four) about 25/30 minutes on their activity, the instructor gets their attention again and discusses the activity. Here is an example of a short one we have then do as they learn about the sample standard deviation and the empirical rule: (Note: we have already done some things with confidence intervals):
Activity #013

Age Measurements for the Shroud of Turin

The Shroud of Turin is a linen fabric that, since 1354, has been claimed to be the burial garment of Jesus Christ. In efforts to establish its authenticity, there has been an enormous amount of scientific testing performed on this object. In one study, several small strips were sent to labs in order to perform radiocarbon dating, a process by which the age of items can be estimated (with some degree of uncertainty). Four of the strips were sent to a lab in Arizona in 1988, resulting in the following estimates for their date of origin: 1397, 1298, 1382, 1287 AD, and their dated age in 1988

<table>
<thead>
<tr>
<th>date of origin</th>
<th>age in 1988</th>
<th>((x - \bar{x}))</th>
<th>((x - \bar{x})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1397</td>
<td>591</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1298</td>
<td>690</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1382</td>
<td>606</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1287</td>
<td>701</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Sigma x =)</td>
<td></td>
<td>(\Sigma(x - \bar{x})^2 =)</td>
<td></td>
</tr>
</tbody>
</table>

1. Use the above table to guide you through the hand calculation of standard deviation.
2. If you go three standard deviations below the mean year of origin, does it seem likely that the fabric could be from the time of Jesus Christ's death?
3. In Minitab, double-check the standard deviation that you calculated. To do this, enter the four data points in a column, and then obtain the descriptive statistics (as you've done before).
4. Now find a 99% confidence interval for the population mean year of origin for the strips sent to the Arizona lab. Go to Stat > Basic Statistics > 1-Sample t. In the "Variables" area, select the column into which you entered the data, and then make sure that the confidence interval level is 99. Interpret the range you obtain. What do the results suggest about the authenticity of the shroud (based on the strips sent to the Arizona lab)?

If you think it would probably have been a good idea to send strips to other labs for dating, you are right -- this was actually done. If you are interested in these and other data and information, the Web has many sites dedicated to studying the Shroud of Turin, such as The Shroud of Turin Website, The Shroud of Turin: Genuine artifact or manufactured relic? (by Jack Kilmon), and The Council for Study of the Shroud of Turin. You may find the controversial debates on these sites to be interesting.
In the last five minutes of the class students take a short lab quiz on-line. There are usually 5 or 6 multiple choice items on the quiz and the students may consult with their lab partner or all of their group members (three or four per group) in answering the questions. The purpose of this, of course, is to encourage 'students teaching students'.

Technology-wise, we use PC's, link to the Internet, use the Minitab Statistical Computing Software, and on-line quizzing software called 'TestPilot'. Student responses on the lab quizzes are sent directly to a file, the results are summarized and made available to instructors to assess student understanding of the concepts covered in the lab. At the next class meeting instructors review any concepts that students did not grasp well. This is great feedback!

3. Assessment

We use a variety of assessments, as follows:

<table>
<thead>
<tr>
<th>Assessment Type</th>
<th>Points</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. In-class exams (2--100 points each)</td>
<td>200</td>
<td>20%</td>
</tr>
<tr>
<td>b. Final exam</td>
<td>200</td>
<td>20%</td>
</tr>
<tr>
<td>c. Readiness Assessment Tests: Individual</td>
<td>160</td>
<td>16%</td>
</tr>
<tr>
<td></td>
<td>Group</td>
<td>80</td>
</tr>
<tr>
<td>d. Homework (best 10 out of 12)</td>
<td>100</td>
<td>10%</td>
</tr>
<tr>
<td>e. Projects (2--one for 60 points, one for 80 points)</td>
<td>140</td>
<td>14%</td>
</tr>
<tr>
<td>f. Lab Quizzes (best 15 out of 19 or 20)</td>
<td>120</td>
<td>12%</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>100%</td>
</tr>
</tbody>
</table>

a. The two in-class exams are 50 minutes in length-- 'open-ended questions' count for about 1/3 of the grade and multiple choice items about 2/3. GTAs grade the open-ended questions and the multiple choice items are machine-graded--answers are placed on bubble sheets (Scantron Forms). The multiple choice portion is graded immediately (usually same day as the exam), scores emailed to students, and an item analysis performed showing (i) student responses, (ii) % correct for each item, (iii) a test validity measure, and (iv) average and standard deviation for all sections of the class. Instructors use the item analysis as an assessment instrument for identifying concepts that students did not grasp well.

b. The final exam is usually entirely multiple choice because of a requirement to assemble and report grades to the registrar's office in a very short time period (two days or less). If the exam is scheduled early in the finals week, a portion of the exam (up to 50%) may be open-ended.

c. Readiness Assessment Tests (RATs) have two (or three) major components, depending on an instructor's use of them: (i) an individual component, (ii) a group component, and (iii) an
appeal process. They were developed by Larry Michaelson [3] as an instructional and assessment tool. Students are given reading assignments before classes and prior to instruction on the material. The goal of the reading is for students to learn some of the basic concepts of the course on their own. After the reading assignments students come to class and take a RAT, made up of true/false and multiple choice questions. These questions should test knowledge and understanding of general concepts and principles rather than small, detailed facts. The goal of the individual RAT is to ensure accountability and understanding of the reading assignments. RATs are given usually given in the LGMs. Students take the individual RAT first and turn it in and then immediately re-take the same test as a group (previously set up) of three to five. The goal of the group RAT is for students to help one another comprehend ideas that they may not have gotten on their own. If the instructor chooses to do so, students are allowed to appeal any incorrect answers based on the quality of the question or a justification for their answer choice. Each student receives an individual and group grade for each RAT.

The instructor uses the feedback from the individual and group RAT scores to determine where students still have misconceptions or misunderstandings. The concepts that students did not get on their own can be used to guide and inform instruction. The feedback helps the instructor focus instruction and activities on application of the course content rather than spending time covering concepts students can easily obtain through self-directed reading and learning. Course activities are typically completed in pairs or groups. The RATs and the content covered on them are used as a means to prepare students for the application of the content in problem-based activities.

RATs cover 'natural units', usually one or two chapters in the text. In a given semester about six or seven RATs are given. RATs provide a powerful motivator for students to read material prior to classes (since it is a major component of their grade) and to keep up with work on a regular basis rather than trying to study at the last minute before an exam.

d. Instructors normally assign 12 homework assignments to be done weekly along with their reading assignments. Some of the homework is discussed in labs or in the LGM. Undergraduate students are hired to grade them, under the supervision of GTAs.

e. Students are given two substantive projects, the first about the fourth week of the semester and the second during the last two weeks of the course. The first is moderately well-structured and the second is fairly unstructured, with just general guidelines. A survey is developed especially for the second project, cooperatively between the students and me. Students are given a scenario, like 'A President of a music company has experienced a
downturn in his business. He asked the marketing department to collect data on aspects related to music, and the Director of Marketing contacted Penn State to do a survey of college students for them. Naturally the Department of Statistics was contacted to do this. So, with the help of the students a survey is created. As a group project, students are asked to analyze the data, interpret it, and prepare a report for the Director of Marketing at the company. The requirements imposed on the project specify that students are to perform at least six different statistical techniques in their analysis (e.g., compare two means, two proportions, regression, chi-square test, analysis of variance, etc.).

f. Lab quizzes were discussed above.

6. Impact on Student Learning

We developed several assessment instruments prior to restructuring the course:

- A content knowledge test consisting of 18 items was developed prior to the restructuring (in 1998/99). It was administered at the beginning and end of the spring 2000, fall 2000 and spring 2001 semesters. During the spring 2000, two sections were taught in the traditional format (n=340) while one section was taught as a pilot using the revised Stat 200. In the fall and spring semesters of 2000/01 all classes were taught using the new format.

- A 20-item test on ‘choosing the appropriate statistical technique from a set of 10 was created in 1996/7 and has been used as part of final exams every semester since then.

- Statistics on D's, F's, gpa's and dropouts were compiled for the 5-year period 1996/97 through 2000/01

- Assessment of student performance in follow-up courses using a subset of the content knowledge test is on-going.

The results:

- The pilot and redesigned classes outperformed the traditional class on the final test of content mastery by 10% to 13% (60%: traditional class, 66% in the pilot class, 68% in the redesigned classes). The improvement in performance in the redesigned class was greatest on concepts. On technical aspects (working with formulas and reading tables for example) the traditional class performed marginally better.

- Students in the restructured course were able to identify the correct statistical technique to use about 86.5% of the time, about 11% better than the 78% correct rate for students in the traditional course. This is viewed as a consequence of lab work.

- The percentage of students receiving a D, F, or dropped the course decreased from a rate of about 12% in the traditional course to about 9.8% in the restructured course, a
decrease of about 18%. The average gpa was essentially unchanged: 2.974 in the traditional course and 3.015 in the new course.

Follow-up data on the performance of students in subsequent courses is too preliminary to draw any conclusions. We will continue to monitor follow-up performance to see if the pattern shown above will be applicable to this as well.

Oh, one more tidbit: our course redesign is saving the department about $115,000 per year!

7. Experiences and Lessons Learned

When we started the restructuring we wanted
i. More hands on involvement with data analysis
ii. Self-Study--More reading and independent learning
iii. Stronger students to help the weaker learn through group work
iv. Students more motivated by lab work than by lecture
v. More contact and communication with students
vi. Less instructor preparation and assessment time needed than previously

How well were our hopes met? What ‘roadblocks’ or disenchantment did we incur?

i. Hands-on Work. This was clearly one of our biggest successes—students definitely had much more hands-on experience with data analysis. Typically students describe, analyze, and interpret 40-50 data-oriented problems during the semester, involving one and two-sample problems about means and proportions, paired comparisons, correlation and regression, analysis of variance, chi-square tests in cross-tabulations, assessing shape of distributions, calculation of normal and binomial probabilities in real-life problems, choosing statistical techniques, and three weeks of project work. On the negative side, we still observe some students just as passive as in large lecture situations.

ii. Self-Study. We hoped students would find ‘self-study’ much more attractive and beneficial than ‘spoon-feeding’ in lectures. This was the case for a sizable proportion of the students—we estimate that 85-90% of the students accepted this approach, albeit grudgingly and stressfully, with the remaining 10-15% quite resistant. We ‘motivated’ them to ‘self-study’ by giving them Readiness Assessment Tests individually and in groups and the high scores in the latter provided sufficient reward for most to accept individual responsibility for learning as a trade-off for lectures. Initially, the stress level was apparently at an unacceptably high level so we backed-off and modified the coverage from 100% new material to somewhere around 35-40% new and 60-
65% previously covered. But this probably resulted in a diminished effort on the part of many students to learn the concepts on their own, mitigating the advantages of the basic approach.

iii. **Stronger students teaching the weaker.** We hoped that weaker students would learn from stronger students and that working in groups would enable better solutions to problems than those obtained by individuals. Our experience on this has been mixed. In general, group solutions are better than individual ones. But weak students get answers from strong students without really learning from them. Group tests have not achieved the purposes for which they were intended—to provide a mechanism for discussion about concepts to help students learn them. Instead, it appears that the primary purpose for the students has been simply to reach an agreement or consensus on answers to questions—essentially but taking a vote.

iv. **Student motivation/sense of value of Statistics.** We hoped that activities and projects would make the course far more interesting, lead to higher motivational levels, and overall result in enhanced performance by the students. As noted earlier we have seen improved performance on understanding of concepts, choosing appropriate statistical techniques, and fewer D’s, F’s, and drops, so that there is solid information that the change in pedagogy has been effective. The change in motivation has not been what we had hoped for, however. But perhaps we are butting our heads against a wall on this: students aren’t far into their majors so it is hard for them to see the need for statistics, regardless of pedagogical method. We are concerned that we have the right system but the wrong course. Perhaps younger students (most are 2nd through 5th semester standing) are not ready—do they need to know more about the role of research in their discipline to be fully ready for this type of approach?

v. **Instructor/Student Interaction.** We hoped that the lab-based system would provide direct interaction between students and instructors—this has definitely occurred and represents a substantial improvement to the situation that existed previously in our large enrolled classes.

vii. **Sharing Instructional Resources.** Before we restructured the course, each instructor prepared their own lecture materials, homework assignments, syllabi, exams, projects etc. Grades were not necessarily comparable because of different criteria. With the new structure, there is substantial uniformity: common syllabus, in-class exams, final exam, Readiness Assessment Tests, and grading standards. We had hoped that we would be sharing development of all course materials and by sharing reduce preparation time greatly. Our experience so far is that preparation time has not decreased and sharing is difficult to forge coordination/agreement on materials, tests, etc. But we are working hard on this aspect and expect a measure of success in the next academic year. An observation: considering how faculty group work goes, it might not be right for us to push students into group work!
Other Comments.

i. Assessment of student understanding of concepts using RATs has proven to be very effective in detecting areas in which students are not grasping the concepts, thereby enabling corrective actions to be taken in a timely manner, and in preparing students for higher level activities in the computer labs than previously., students have been helped in building skills, as the evidence of the pre As a result - and post-test shows. The web page enables more rapid feedback to students, another crucial element in the learning process.

Student perception of the importance of RATs is evident in the results from Innovation and Quality (IQ) survey data where the majority of students (55%) rated the RATs as one of the most important aspects of the class. Seventy-five percent of respondents believed that periodic RATs help them keep up with the readings and that they were vital for their learning and understanding of the content. As voiced in focus groups, students felt that the RATs helped by promoting recognition of holes in their understanding. In addition, students liked the opportunity to work in groups and interact with others in the class. Most students emphatically suggested keeping the Readiness Assessment Testing as part of the course.

ii. As for the individual and group collaborative activities, we were pleasantly surprised at the students' reaction to not being lectured to and instead being able to work in groups in the labs to apply what they had learned from the resource

Penn State University has about 20 other campuses around the state (in addition to the main campus at University Park). We will be exporting a modified version of our model to these campuses in the next academic year.

References.
[6] For information about Test Pilot, see http://www.clearlearning.com