

Designing and Evaluating Assessments for Introductory Statistics

*Robert delMas – University of Minnesota
delma001@umn.edu
Beth Chance – Cal Poly, San Luis Obispo
bchance@calpoly.edu

Workshop W26, November 12, 2005

Schedule (1:15-3:15):

- I. Introductions, Overview of Assessment
- II. Overview of ARTIST Materials
- III. Discussion of Artist Assessment Builder
- IV. Developing Individual Assessment Plans

ARTIST (Assessment Resource Tools for Improving Statistical Thinking) Website:

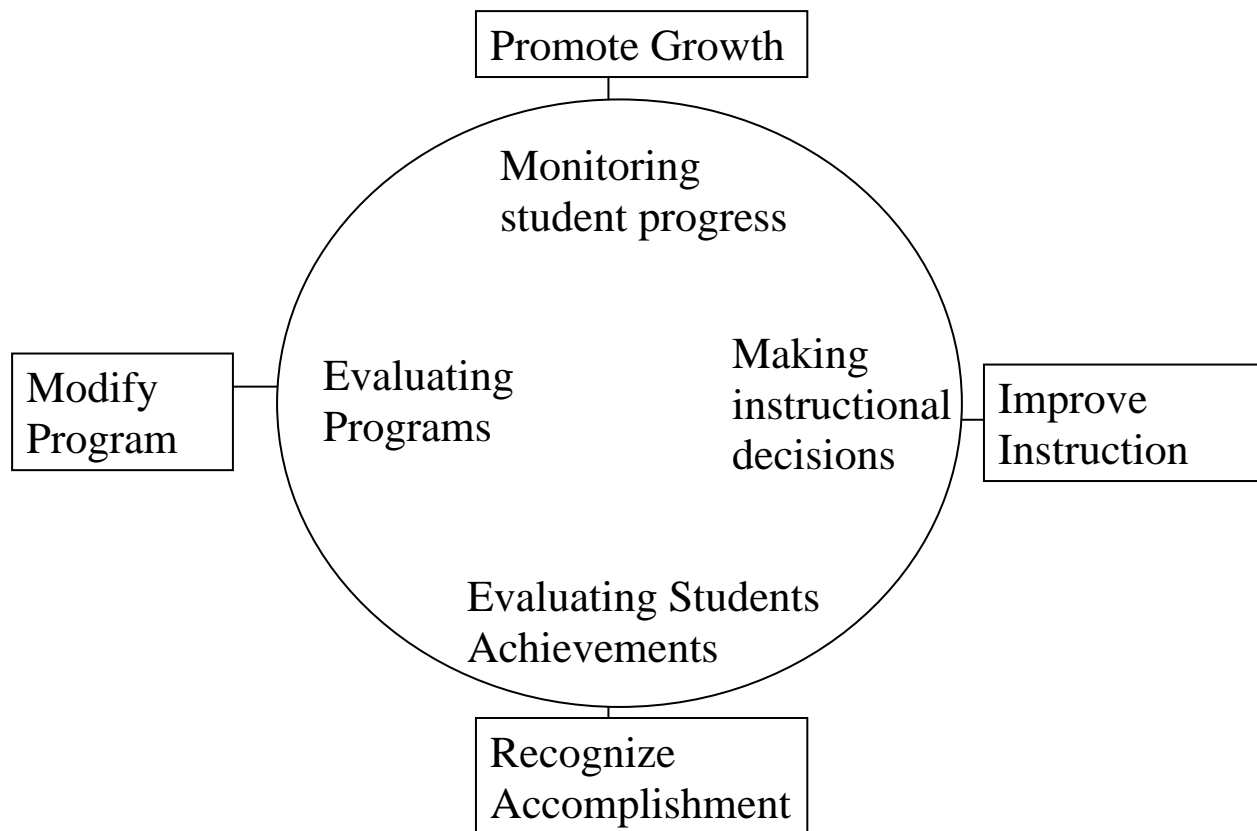
<http://ore.gen.umn.edu/artist/>

PI: Joan Garfield, University of Minnesota

Assessment = on-going process of collecting and analyzing information relative to some objective or goal.

Evaluation = interpretation of evidence, make a judgment (quality, worthiness, appropriateness, goodness, validity, etc.), comparisons between what was intended (learning, progress, behavior) and what was obtained, use information to make improvements.

Dimensions of Assessment



Types of Assessment

Formative = In-process monitoring of on-going efforts in attempt to make rapid adjustments

Summative= Record impact, compare outcomes to goals, decide next steps

Learning Skills Program

Bloom's Taxonomy *

Benjamin Bloom created this taxonomy for categorizing level of abstraction of questions that commonly occur in educational settings. The taxonomy provides a useful structure in which to categorize test questions, since professors will characteristically ask questions within particular levels, and if you can determine the levels of questions that will appear on your exams, you will be able to study using appropriate strategies.

Competence	Skills Demonstrated	Implication
Knowledge	<ul style="list-style-type: none"> • observation and recall of information • knowledge of dates, events, places • knowledge of major ideas • mastery of subject matter • <i>Question Cues:</i> list, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc. 	
Comprehension	<ul style="list-style-type: none"> • understanding information • grasp meaning • translate knowledge into new context • interpret facts, compare, contrast • order, group, infer causes • predict consequences • <i>Question Cues:</i> summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend 	
Application	<ul style="list-style-type: none"> • use information • use methods, concepts, theories in new situations • solve problems using required skills or knowledge • <i>Questions Cues:</i> apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover 	

<p>Analysis</p>	<ul style="list-style-type: none"> • seeing patterns • organization of parts • recognition of hidden meanings • identification of components • <i>Question Cues:</i> analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer 	
<p>Synthesis</p>	<ul style="list-style-type: none"> • use old ideas to create new ones • generalize from given facts • relate knowledge from several areas • predict, draw conclusions • <i>Question Cues:</i> combine, integrate, modify, rearrange, substitute, plan, create, design, invent, what if?, compose, formulate, prepare, generalize, rewrite 	
<p>Evaluation</p>	<ul style="list-style-type: none"> • compare and discriminate between ideas • assess value of theories, presentations • make choices based on reasoned argument • verify value of evidence • recognize subjectivity • <i>Question Cues</i> assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize 	

* Adapted from: Bloom, B.S. (Ed.) (1956) Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York ; Toronto: Longmans, Green.

An Assessment Cycle

1. Set goals (concepts, skills, applications, attitudes, beliefs)
2. Select methods
 Purpose: why, how used
 Who: student, peers, teacher
3. Gather evidence
4. Draw inference
5. Take action
6. Re-examine goals and methods

Setting Goals:

What should be learned? What should the student be able to know, do, and understand?
 What do you value? At what point in the course should students develop this knowledge/skills?
 Translate your goals into assessable learning outcomes/objectives

Selecting Methods:

What method(s) of assessment is best suited to students demonstrating achievement of the learning outcomes?

Gathering Evidence:

How will I know how well the student has achieved the learning outcome?

Why do we need to consider other assessment methods?

Formal summative exams are best at assessing knowledge, basic skills such as isolated computational skills, and memory retrieval. They provide consistent and timely scoring (how many right answers?). They predict future performance. They are expected and allow extensive course coverage (especially multiple choice).

Not as effective at assessing background skills, such as prerequisite knowledge, teamwork and communication skills, and broader analytical and problem solving skills. They don't tell us enough about what exactly the students do and do not understand, and what they can and cannot do with their knowledge. They often are divorced from context, focus only on accuracy of computations, correct applications of formulas, correctness of graphs and charts.

There is more to statistics and the learning of statistics, including critical thinking, analytic thinking, interpretation. Many of these skills do not have exclusively "right or wrong" answers.

Examples

1) "What Went Wrong?"

Giving students practice in diagnosing (someone else's) errors in low-stakes environment. Try to focus on most common misconceptions.

(Chance and Rossman, 2005) The following table presents data on the "number of inversions" made by 144 roller coasters during one run, as obtained from the Roller Coaster DataBase (www.rcdb.com).

Number of inversions	0	1	2	3	4	5	6	7
Tally (count)	84	7	12	11	9	9	3	9

Consider the following possible solutions for calculating the mean and median. Identify what the student did incorrectly in each case.

- (a) A fellow student reports the median to be 3.5 inversions.
- (b) A fellow student reports the mean to be 18 inversions.
- (c) A fellow student reports the median to be 9 inversions.
- (d) A fellow student reports the median to be 0. (The problem here is an error of presentation, not calculation.)

2) "Practice Problems"

Giving students routine practice on key concepts and immediate application of current topics. Can also challenge students before next topic and/or basic calculations to free class time to focus more difficult concepts.

(Chance and Rossman, 2006) More on sleep deprivation

(a) Recall the difference in group medians between these two groups.

(b) Change one line of your Minitab macro from Investigation 2-8 so that it analyzes difference in group medians rather than differences in group means. Apply your macro to the original data from the sleep deprivation study, using at least 1000 repetitions. Describe the distribution of differences in group medians, and report the empirical p-value.

(b) Comment on how the randomization distribution of the differences in medians differs from your earlier analysis of means.

(c) What conclusion would you draw about the statistical significance of the observed difference in group medians from this analysis? Explain.

3) Exam Questions

Focus on students' ability to interpret, integrate understand overall process, understanding of fundamental statistical concepts. Opportunity to convey to students what is most important to you.

(a) Your text states that "confidence intervals seek to estimate a population parameter with an interval of values calculated from an observed sample statistic." Convince me that you understand this statement by describing a situation in which one could use a sample proportion to produce a confidence interval as an estimate of a population proportion. Clearly identify the population, sample, parameter, and statistic involved in your example. Do not use any example that appears in your book.

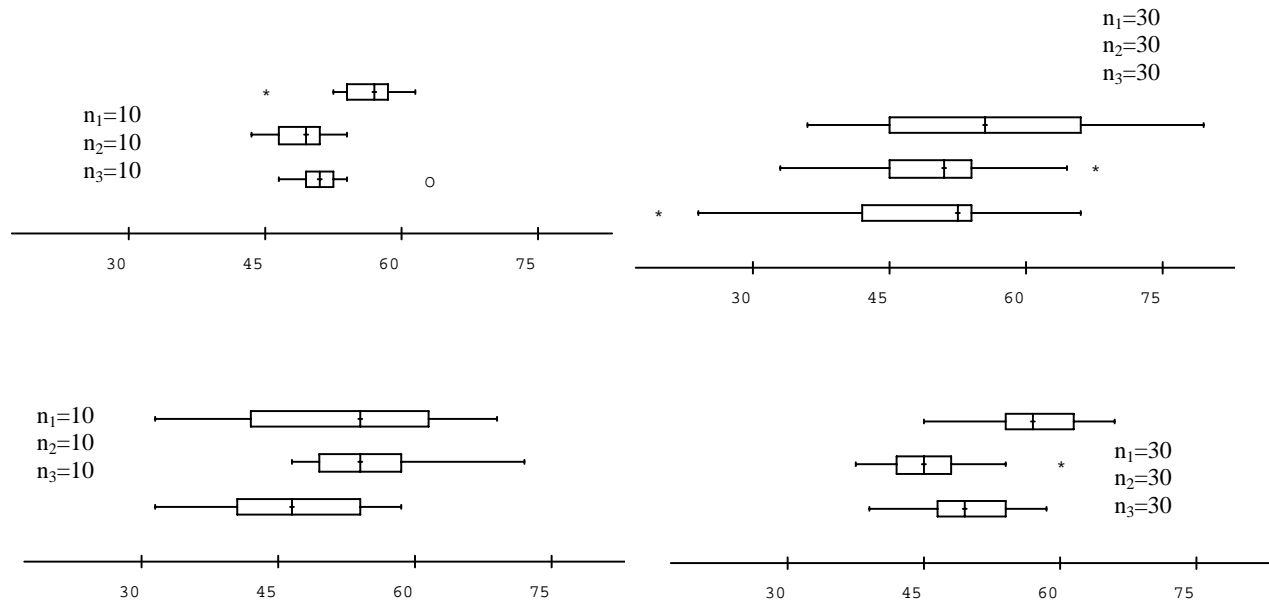
(b) (Scheaffer et. al.) A book review in the *Los Angeles Times* of December 29, 1993 about discrimination against upper-class African-Americans contains the following example:

A Harvard law student says her professor tended to ignore the raised hands of the black students in class-and then, suddenly he would call on several black students in a row: "As if", she explains, "the professor had suddenly realized that he was neglecting an important segment of the student body and had resolved to make amends".

Briefly discuss this quotation in light of what you have learned from the "longest run of heads" activity about common perceptions of streak behavior. Do you agree with her argument?

(c) Four different studies obtained data that were used in a test of the hypothesis $H_0: \mu_1 = \mu_2 = \mu_3$. Based on the information below, order these studies from smallest p-value to largest p-value.

Provide an explanation for your choices. You will be graded primarily on your explanations.



(d) Researchers conducted a “randomized, double-blind trial” to determine whether taking large amount of Vitamins E protects against prostate cancer (*Journal of the National Cancer Institute*, 1998). To study this question, they enrolled 29,133 Finnish men, all smokers between the ages of 50 and 69. The men were randomly divided into two groups: One group took vitamin E and a second group took a placebo. The researchers followed all the men for eight years and then determined how many had developed prostate cancer. They found that participants taking vitamin E were “significantly” less likely to develop prostate cancer.

- Explain what “randomized” means in this study and its purpose.
- Explain what “double-blind” means in the context of this study and its purpose.
- Explain what “significantly less likely” means in a statistical sense and why it is an important consideration.
- Based on this report, would you consider it reasonable to conclude that taking vitamin E causes a reduction in the probability of developing prostate cancer? Explain your reasoning.
- Based on this report, what population would you be willing to generalize these results to? Explain your reasoning.

(e) Statistical evidence played an important role in the murder trial involving Kristen Gilbert, a nurse who was accused of murdering hospital patients by giving them fatal doses of heart stimulant. Hospital records for an eighteen-month period indicated that of 257 eight-hour shifts that Gilbert worked on, a patient died in 40 of those shifts (15.6%). But of 1384 eight-hour shifts that Gilbert did not work on, a patient died in only 34 of those shifts (2.5%).

- Identify the observational units in this study.
- Identify the explanatory variable and the response variable in this study.
- Organize the given information into a two-way table, putting the explanatory variable in columns and the response variable in rows.
- Calculate the odds ratio of a death occurring on a shift, comparing shifts on which Gilbert worked to shifts on which she did not work.
- Treat these data as a random sample from a population, and produce a 95% confidence interval for the population odds ratio.
- Interpret what this confidence interval reveals about the question of whether a significantly higher proportion of deaths occurred on Gilbert’s shifts as compared to other shifts.

g) Put yourself in the role of the defense attorney who needs to argue that Gilbert was not responsible for any deaths. Suggest a potential confounding variable that you might use to explain why there was a higher percentage of deaths on Gilbert's shifts. Explain how this confounding variable provides an alternative explanation to the prosecution's contention that Gilbert was responsible.

(f) It can be shown that the sum of the residuals from a least squares regression line must always equal zero.

a) Does it follow that the mean of the residuals must always equal zero? Explain briefly.

b) Does it follow that the median of the residuals must always equal zero? Explain briefly.

(g) Suppose that every student in this class scores 10 points lower on the final exam than on the first midterm exam. What would be the value of the correlation coefficient between midterm exam score and final exam score? Explain briefly.

Assessment Implementation Issues

The ARTIST website contains a section in which ten advisors to the project provide their own responses to a variety of questions about how they implement assessment of student learning in their own courses.

To access this section from the main ARTIST web page (<https://ore.gen.umn.edu/artist/>), click on Resources (along the left side) and then Implementation Issues (#6 on the list).

The general areas of discussion considered there are:

- *Use of Exams*: How are exams used in the course (format and purpose)?
- *External Aids*: Do student use external reference aids on exams?
- *Use of Technology*: Are students allowed/encouraged to use any technology on exams?
- *Constructing Exams*: How are questions chosen and points assigned?
- *Writing Process*: Describe the process used to write an exam from scratch?
- *Exam Grading*: How are questions graded and scores reported?
- *Preparing Students for the Exam*: What preparation is supplied for the students?
- *Post-Exam Feedback*: How is exam performance conveyed to the students afterwards?

Holistic Scoring Rubric

- 5 points: discussion of effects of sample size, differences in means, and variability
- 4 points: ignores sample size
- 3 points: only focuses on differences in means
- 2 points: ordering with no explanation

As the number of points on the scale increases, so does the difficulty in developing and applying the scale. Often best to start with 4 points until get more comfortable.

Other suggestions (see Hubbard; Wild, Triggs, & Pfannkuch):

- Multiple choice
 - with identification of false response
 - with explanation or reasoning choices
 - with judgment, critique (when is this appropriate)

- “What if”, working backwards, “construct a situation such that”
- Objective-format questions
 - e.g., comparative judgment of strength of relationship
 - e.g., matching boxplot with normal prob plots
- Missing pieces of output, background

Sample Final Exam

1) The following data are the point totals for the Men’s Basketball team in their first 8 victories this season:

80 72 68 55 80 78 90 85

- (a) (5 pts) Make a stemplot of these winning point totals and describe the shape of the distribution.
- (b) (3 pts) Would the Five Number Summary or the mean and standard deviation be a better summary for this distribution? Explain your choice.

2) Two investigators wanted to study the heights of 18-24 year old men in Stockton. One investigator, Happy Harry, took a random sample of 100 men. The other investigator, Tired Tony, took a random sample of 1000 men.

- (a) (2 pts) If each investigator finds the average height of the men in his sample, which investigator, Harry or Tony, should expect a larger average, or should they be about the same? Explain.
- (b) (3 pts) Which sample, Harry or Tony’s, do you expect will have less bias or will they be about the same? Explain.
- (c) (3 pts) Which estimate of the population mean, Harry or Tony’s, should have higher precision, or will they be about the same? Explain.

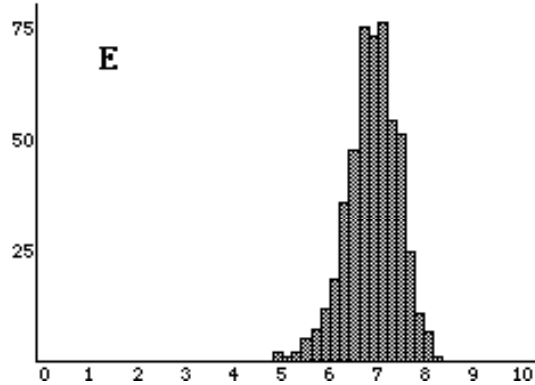
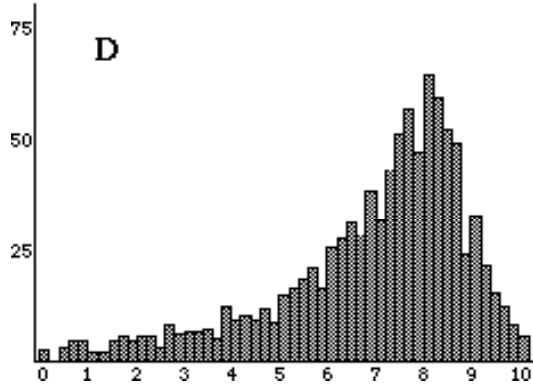
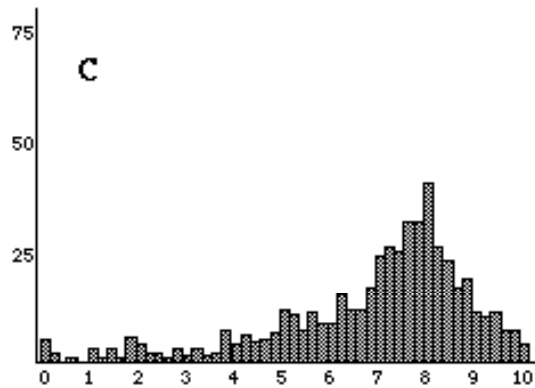
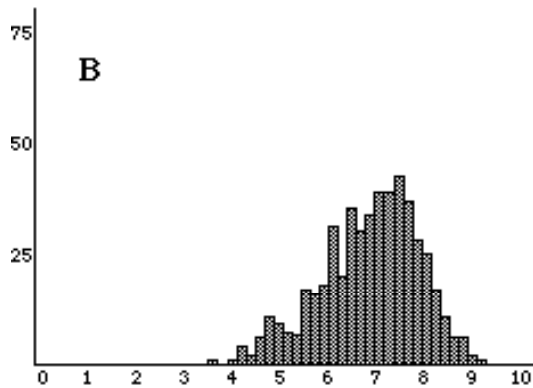
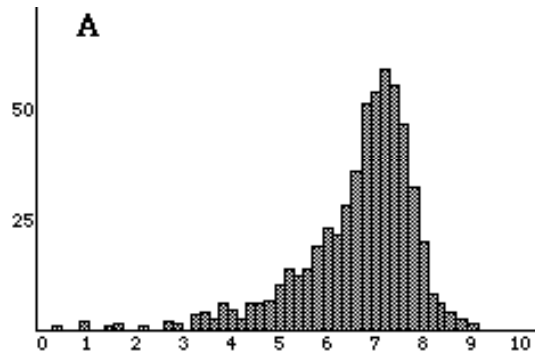
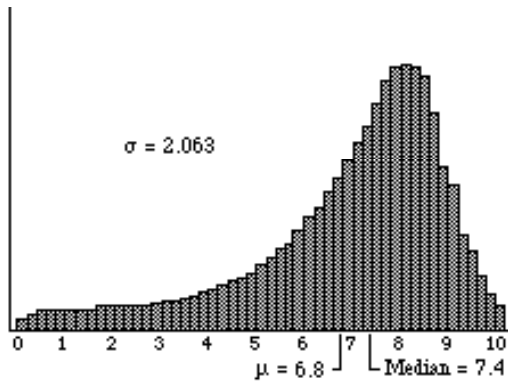
3) In 1988, men averaged about 500 on the math SAT, the standard deviation was about 100, and their scores followed a Normal distribution. One of the men who took the math SAT in 1988 will be picked at random, and you have to guess his test score. You will be given 50 dollars if you guess it right to within 50 points.

- (a) (2 pts) What *one number* should you guess?
 - (b) (5 pts) With this guess, what is your probability of winning the 50 dollars? Explain.
- Extra Credit: What is your expected winnings?

4) The distribution for a population of test scores is displayed below on the left. Each of the other five graphs, labeled A to E represent possible sampling distributions of sample means for 500 random samples drawn from the population. (*Justify choices*)

- (a) (2 pts) Which graph represents a sampling distribution of sample means for samples of size 1?
 A B C D E
- (b) (2 pts) Which graph represents a sampling distribution of sample means for samples of size 9?
 A B C D E

Population Distribution



5) A social research scientist wants to test whether the percentage of Republicans who favor the death penalty is greater than the percentage of Democrats who are in favor of the death penalty. Suppose the sample data showed that the percentage of Republicans who are in favor of the death penalty is 42% and the percentage of Democrats who are in favor of the death penalty is 40%.

- (a) (2 pts) Write down the null and alternative hypotheses for this test.
- (b) (3 pts) The p-value for this test is .0021. The 95% confidence interval for $p_1 - p_2$ is (.00637, .03363). Which of the following conclusions do you think is more appropriate to draw?
 1. There is evidence of a large difference in the two proportions.
 2. There is strong evidence of a difference in the two proportions.

Explain.

(c) (2 pts) Which conclusion does a p-value better address? Explain.

(d) (2 pts) Which conclusion does a confidence interval better address? Explain.

6) In a clinical trial, data collection usually starts at “baseline,” when the subjects are recruited into the trial but before they are randomized to treatment and control groups. Data collection continues until the end of follow-up. Two clinical trials on prevention of heart attacks report baseline data on weight, shown below.

		Number of persons	Average weight	Standard deviation
Trial 1	Treatment	1,012	185 lb	25 lb
	Control	997	143 lb	26 lb
Trial 2	Treatment	995	166 lb	27 lb
	Control	1,017	163 lb	25 lb

(a) (4 pts) In one of these trials, the randomization did not achieve the desired result. Which trial and why do you say so? How will this affect our results and conclusions for this study? (Hint: make sure you focus on the most serious difficulty)

(b) (4 pts) Below are ten people and their weights. Randomly assign them to one treatment group and a control group (start with line 139 of the random number table). Clearly show your work.

Bob 148 Tom 174 Joe 148 Fred 133 Sam 157
 Curt 177 Al 162 Harry 188 Gami 160 Dan 188

7) Can pleasant aromas help a student learn better? Two researchers believed that the presence of a floral scent could improve a person’s learning ability in certain situations. They had ten people work through a pencil and paper maze 2 times, first wearing an unscented mask and then wearing a scented mask. Tests measured the length of time it took subjects to complete each of the two trials. They reported that, on average, subjects wearing the floral-scented mask completed the maze more quickly than those wearing the unscented mask.

(a) (3 pts) Is this an observational study, survey, or experiment? Explain.

(b) (2 pts) Identify the response and explanatory variables.

(c) (4 pts) Explain how confounding makes the results of this study worthless.

(d) (4 pts) Sketch an outline of a more appropriate design for the study.

8) NCAA collected data on graduation rates of athletes in Division I in the mid-1980s. Among 2,332 men, 1,343 had not graduated from college, and among 959 women, 441 had not graduated.

(a) (3 pts) Set up a two-way table to examine the relationship between gender and graduation.

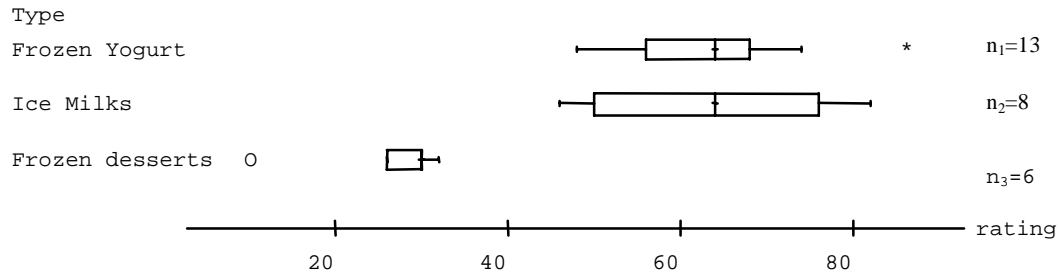
(b) (3 pts) Calculate a couple of *conditional* percentages to describe the relationship between gender and graduation.

(c) (3 pts) Identify a test procedure would be appropriate for analyzing this relationship? State the null and alternative hypotheses.

(d) (3 pts) What type of distribution does the test statistic you describe in (c) follow? For what values of this test statistic will you reject the null hypothesis at the 5% level?

(e) (2 pts) If the above result is significant, would this mean that if some people have a sex change they will increase their chance of graduating? Explain briefly.

9) A panel of trained testers judged the flavor quality of different vanilla frozen desserts (frozen yogurts, ice milks, other frozen desserts) measured on a scale from 0 to 100. The data are from a *Consumer Reports* article “Low-fat frozen desserts: Better for you than ice cream?” (August, 1992). Below is a graphical summary of the data.



Here is most of the ANOVA output from the computer:

ANALYSIS OF VARIANCE ON rating

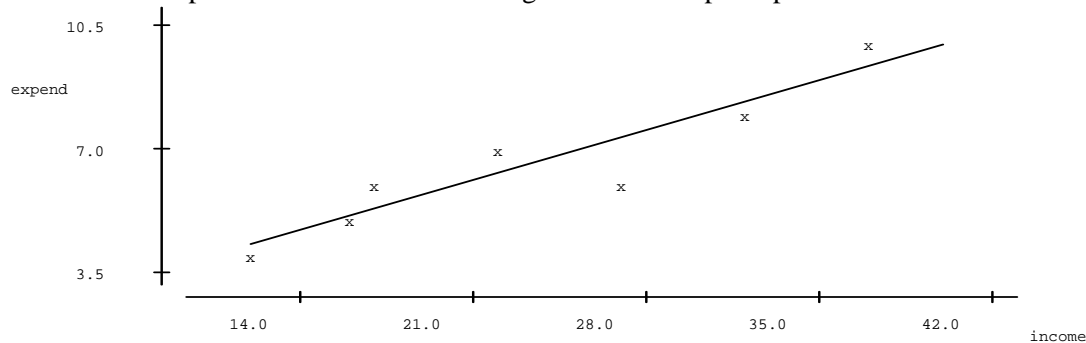
SOURCE	DF	SS	MS	F	p
TYPE		6364	3182		
ERROR	24	3031	126		
TOTAL		9395			

- (2 pts) Explain briefly why ANOVA was the appropriate analysis for these data.
- (2 pts) State the null and alternative hypotheses.
- (4 pts) Finish the ANOVA table giving the F-statistic, degrees of freedom, and approximating the p-value. Show your work. What is your conclusion about the flavor quality of the different desserts?
- (2 pts) Based on the graph, do you feel the technical conditions needed for the validity of this test procedure are valid? Explain.

10) A random sample of 7 households was obtained, and information on their income and food expenditures for the past month was collected. The data (in hundreds of dollars) are given below.

Income (\$100's)	22	32	16	37	12	27	17
Food Expend (\$100's)	7	8	5	10	4	6	6

Here's a scatterplot of these data with the regression line superimposed.



Here's the Minitab output:

The regression equation is $\text{expend} = 1.87 + 0.202 \text{ income}$

Predictor	Coef	Stdev	t-ratio	p
Constant	1.8690	0.9068	2.06	0.094
income	0.20195	0.03661	5.52	0.003

$s = 0.8181$ $R\text{-sq} = 85.9\%$ $R\text{-sq}(\text{adj}) = 83.1\%$

- (2 pts) Describe the direction and strength of the association.
- (2 pts) On the graph, identify the point which you think has the largest residual. Explain.
- (2 pts) On the graph, identify the point which you think has the most influence on the position of the regression line, and how the line would change if it was removed. Explain.
- (3 pts) Provide an interpretation of the number .202 in the regression equation in the context of these data. Exactly what does this value tell us?

- (e) (4 pts) Is there evidence of a statistically significant relationship between income and food expenditure? Make sure you clearly explain the basis for your answer.
- (f) (2 pts) Explain why you would not recommend using this relationship to predict the food expenditure for a household with an income of \$5,200.

11) National data show that, on the average, college freshmen spend 7.5 hours a week going to parties. President DeRosa doesn't believe that these figures apply at UOP. He takes a simple random sample of 50 freshmen, and interviews them. He finds that the 95% confidence interval for the mean number of hours spent a week going to parties is (5.72, 7.42).

(a) (4 pts) Explain to the President what he means by the phrase "95% confidence."

Now he wants to test the hypothesis that the mean for UOP is different from the national mean at a 5% significance level.

(b) (2 pts) Specify the null and alternative hypotheses for this test.

(c) (2 pts) Indicate a test procedure he could use to conduct this test.

(d) (3 pts) Eager to gain favor with the president, you tell him that you can save him lots of time because, based on the data already presented, you know what he will conclude and he doesn't have to perform any additional calculations. Does he reject or fail to reject the null hypothesis at the 5% level? Explain.

Extra Credit

Suppose you take 50 measurements on the speed of cars on Interstate 5, and that these measurements follow roughly a Normal distribution. Do you expect the standard deviation of these 50 measurements to be about 1 mph, 5 mph, 10 mph, or 20 mph? Explain.

Methods of Assessment (*Assessment Practices in Undergraduate Mathematics*, MAA Notes 49)

- *Timed, in-class, individual efforts*
 - Watch balance of speed vs. skill, inter-dependencies of subsequent subparts
 - Watch balance of conceptual vs. calculation questions
 - Consider richer types of questions versus mimicking examples
 - Consider holistic rubrics (can discuss ranking of quality of responses with students)
 - Discuss results with students afterwards (problem-solving strategies, common misconceptions)
 - Watch balance with other assessment tasks in the course
 - Consider requiring/evaluating students to write questions
 - Consider using multiple-choice as formative assessments, focusing on the underlying misconceptions
 - Can ask students to rate their level of confidence on a question
- *Written homework assignments/lab assignments*
 - Most appropriate for application of methods discussed in class
 - Require explanation/demonstration of steps for full credit (more important than final answer)
 - Consider including more conceptual questions (instructor can grade only these?)
 - Consider alternative weighting schemes of assignments
 - Consider asking/requiring students to reflect (significantly) on the problem solving process
 - Provide clear guidelines of expectations, help students avoid time consuming sidetracks
 - Provide feedback for future assignments
 - Consider encouraging questions on assignments with Socratic responses
 - Consider "charging" students for responses...

Consider limiting when questions can be asked (not night before)/encourage email questions
Consider group assignments with group grade
Provide access to model solutions (after the fact), can even be anonymous student paper

- *Minute Papers*

Provide very quick assessment of student comfort, understanding, key concerns
Hear from *all* students (recommend anonymous with option for individual feedback)
What we think we said versus what they heard
Establish student comfort with critical comments, clarify what you want
Consider sharing common responses with students afterwards, be considerate
 Students appreciate opportunity to give input, knowing they are not alone in their confusion/concerns
Open-ended or directive (most important point(s), unanswered questions, learn more about?
Muddiest point? what has worked best in how this class is taught, what has not worked?)
 Develop skill to identify important points/ask good questions
Consider giving bonus credit

- *Expository writings (e.g., journals, essays, on homeworks/exams)*

Focus on student ability to reflect, explain, integrate, use the language of statistics
Expect/require improvements in content and composition as course progresses (make sure move beyond “surface” level writing)
Consider varying the “audience” of the writing assignment (e.g., soda manufacturers, manager of baseball team, other students, younger sibling)
Explain to students the *why*, the intentions
Recommend requiring word-processing
Provide students with sufficient guidelines, details of your expectations (e.g., model papers)
Use scoring rubric (need to be better at giving more global feedback on the writing structure, style), provide in advance (consider requiring to attach), especially on first paper, aids in appearance of objectivity. Consider asking students for self-evaluation using rubric.
Provide guidance for future assignments
Consider mini-deadlines and option for rework (tape recorded feedback?)

- *Portfolios/Journals*

Students asked to choose and collect work of high quality over time. Takes some pressure off individual assignments.
Reflective portfolios focus on student growth over time. Provide insight to how understanding develops over the course, particularly difficult areas, and student attitudes.
Can ask students to keep a *journal* of their reflections, difficulties, feelings during the course.
 Maybe want to leave ungraded or consider part of class participation, helps open dialog between student and professor. Provide clear guidelines.
 May want to focus some of the topics, what do you want to learn?
Project portfolios focus on collection of student project work, paralleling a professional portfolio collection, e.g, title page, abstract, references, acknowledgements.
Can include requiring students to submit questions on the reading (early in week) or summarizes of the chapter (end of week).

- *Student projects*

Assess student ability to apply statistical *process*, help relate material to real world

Students must select and justify relevant tools for a question of interest (e.g., their choice)

For term-long projects, incorporate mini-deadlines with opportunity for feedback

Provide scoring rubric in advance (aids consistency in grading), even day 1! Review during course.

Consider providing access to model papers/presentations/proposals

Consider peer-review of proposals (provide rubric), can be anonymous

Consider having students log hours spent on project/provide feedback on process, warn may be individual grade adjustments (e.g., 75% group, 25% individual). Consider follow-up quiz or exam questions.

Consider assigning and/or rotating roles (e.g., convenor, recorder, organizer), especially if require reflection on group process as part of write-up.
- *Paired quizzes/group exams*

Significant student learning occurs in the debates they have with each other. Students also help and encourage each other. Helps alleviate early anxieties. Can challenge further.

Become much more of a learning activity themselves.

Group dynamics seem to work better than with group projects?

Have them check each others' work

Discuss the plan with them early in course to motivate fully participation

Give individuals time to think quietly before talking with other group members

Include additional individual components (maybe weight more heavily as course progresses)

Helpful to rearrange groups frequently. Experiment with different grouping approaches.

Aim for 4 people/group.

Give specific objectives and timelines to keep the group on task.

Can group together those who don't want to work in groups

Consider discussing effective strategies with them (e.g., listen to each other's opinions openly, support solutions that seem logical and objective, avoid tricks such as majority vote to reduce conflict).

Consider supplying questions ahead of time to group and then interviewing members individually during an oral presentation.

Can allow discussion but then submission of individual work. Can have them take the exam twice, once alone (submit work) and once together, can average scores.
- *Concept Maps*

Strengthen students' understanding of how a new concept is related to other (known) concepts.

Can provide preset schematics and lists of terms for them to fill in

Need to be careful not to give away too much too quickly for them

Can combine with student review of text and small group discussions

Focus on content and relationships between concepts

Be open to unexpected, creative responses

Especially helpful for learning what needs review, what connections students are not making among theories/concepts, how students are organizing their thoughts, what

preconceptions/prior knowledge is present prior to introducing a topic, what changes occur in conceptual representations.

Helpful for more visual, less verbal students.

Formal Writing Assignments

Give students authentic tasks in which to apply their statistical knowledge in layman's terms.

Examples

1) Joy Jordan (see ARTIST Roundtable Proceedings)

Assignment #1

Suppose you receive the following letter from your dad:

Hey Kiddo,

I am worried about Grandma. Remember that she was diagnosed with high blood pressure? Well, she's currently taking the medication Makemewell to lower her blood pressure. At the time of Grandma's diagnosis, her doctor said that a randomized, double-blind experiment had been conducted and that Makemewell was shown more effective in lowering blood pressure than a placebo. To be honest, I have no idea what any of that means, but I believed and trusted the doctor. Now I've heard two stories that make me think differently. Larry, our next-door-neighbor, was taking Makemewell and he got a terrible fever that put him in the hospital. Also, my coworker, Sally, actually had her blood pressure go up while she was taking Makemewell! I'm now very suspicious of this medication.

I know that you're taking a statistics course at college. Based on the information I've given you, do you think Grandma should stop taking her medication? Whatever your opinion, will you please explain yourself thoroughly and clearly? (I will draw on your responses when I talk with the doctor.) And please don't use any statistics mumbojumbo that I won't understand. I really appreciate your help with this.

Love, Dad

Your assignment is to type a 1-2 page letter (single-spaced, 12-pt. font) responding to your dad.

Grading Criteria (30 points possible)

- ___ The explanation to your dad convinces me (your teacher) that you understand the statistical concepts involved in the assignment. (14 points)
- ___ The explanation to your dad is thorough, well organized, and clear. (6 points)
- ___ The explanation to your dad is presented in non-technical terms that he will understand. (5 points)
- ___ You successfully paid attention to accepted conventions of language use (syntax, spelling, grammar, readability, etc.) (5 points)

Sampling scoring rubric of essay: (Crannell, 1999)

Does this paper:

1. clearly (re)state the problem to be solved?
2. state the answer in a complete sentence which stands on its own?
3. clearly state the assumptions which underlie the formulas?
4. provide a paragraph which explains how the problem will be approached?
5. clearly label diagrams, tables, graphs, or other visual representations of the math (if these are indeed used)?
6. define all variables used?
7. explain how each formula is derived, or where it can be found?
8. give acknowledgment where it is due?

In this paper,

9. are the spelling, grammar, and punctuation correct?
10. is the mathematics correct?

11. did the writer solve the question that was originally asked?

Total number of yeses = grade on paper.

2) Nathan Wetzel (see ARTIST Roundtable Proceedings)

Class Data Collection and Analysis – Assessment

Assignment I

(10 points) We want to design a study of fast food. We have an imaginary client who has asked us to get answers to the following questions.

1. How many ounces are in the average McDonald's Large French Fry? Small French Fry?
2. How often does the average UWSP student eat fast food? Is there a difference between men and women?

If you have any questions for our client, Professor Wetzel will play the role of the client.

Describe a practical, reasonable, usable way for our class to get data that will help in our investigation. DO NOT collect any data. Your description should include:

1. a restatement and clarification of the investigation.
2. For the French fry part, what measurements(s) to make and how to make them. (we need a way to get the data which allows everyone to make some measurements – I agree that ideally, we would have the same person make all of the measurements, but I want everyone measuring.)
3. For the fast food part what questions would you ask students? Include specifics on wording and format.
4. at least one additional demographic question (like gender) for which we could compare fast food use.
5. how to get random samples.
6. any other data that we might want to collect.

We will have 50 students helping collect the data. We will distribute the work equally among these 50. Also, this data collection will cost us all a small amount of money. Arrange your data collection so that we use no more than 100 orders of French fries.

This problem requires thinking and possibly some research. A good answer will get 7 out of 10 points, in order to get full credit, you need to provide a great answer.

We will discuss your answers in class and later we will collect data using our consensus answer. An Example of a bad answer to this question: (this answer would get 2 points - minimal restatement, no specifics, lacks explanation, no extras)

"We are investigating French fries and fast food. I would have everyone go to McDonalds and buy a order of French fries. Then we would weigh it. We would also ask students how often they go to a fast food restaurant."

Rubric for grading this Problem

(5 points) Basics: Does the description include all of the required parts including accurate

and appropriate use of terminology?

- restatement
- measurements - including specifics
- questions to ask - including specifics
- extra demographic question
- how to get randomize - including specifics
- other data

(2 points) Organization: Is the description organized and neatly presented?

Great answers also include some pluses.

Pluses:

- Does the description include any significant extras?
- Significant improvements to a basic data collection design.
- Extra thought into the specifics of this context - recognizing potential problems and giving solutions.

Minuses:

- Does the description indicate that the student is mimicking a book answer and not considering the context?
- Does the description include a design that would be extremely impractical?
- Does the randomization described introduce a significant confounding variable that was not identified?

Second Assignment:

(22 points) Analyze all interesting portions of the data and write a short report (at least 2 pages, double spaced, typed, not including graphs) to our client summarizing our data collection, your analysis and your interpretation. Write the report to a client who understands p-values, but is not interested in the details of the computations. Your client understands all of the graphical methods that we have looked at. In other words, don't include MINITAB hypothesis test output, but do include graphical summaries.

In addition, include a short letter that we could send to McDonalds. Notice that this part is the main part of the assignment. 'Good' answers will receive 14 points, you must do a great job to receive full credit. In other words, I want YOU to ANALYZE the data. This part is open ended and in order to do a great job, you will need to look at MINITAB output that I did not explicitly tell you to get. You will also need to THINK AND CARE about the data. Answer the questions that the client should have asked or that you saw interesting results.

Goals of Assessment:

1. The students know from the beginning that this problem will be assessed differently. I have found that many are very satisfied to get a score of 70% because they knew that their solution was average.
2. The students have access to solution to similar problems. Every semester, I choose a different topic, so solutions from past semesters do not give it all away." The sample "great" solution is fantastic, so it sets a high standard. This helps with the students who want to do exactly enough work to get an "A" and no more.
3. The assessment is done and the problem is handed back in a short time period. This helps keep the topic current for the students.
4. The students who did not receive full credit have a chance to see what kinds of solution did receive full credit. The timing of this often serves as a review of the chapter, but it also automatically answers most of the questions that the "point hungry" students ask and encourages more questions from the "knowledge hungry" students.

Projects

Involves students in the entire statistical process, messiness of real data, focus on communication and even collaboration

Examples

1. B. Chance – term long project

Goal: To collect, describe, and analyze data to answer two questions of your choice. This will allow you to apply the skills you learn in this course to the world around you which will in turn enhance your appreciation and retention of the material.

Teams: For the class projects, you will work in groups of 1-4. It is up to the members of the group to make sure everyone contributes equally. Teams should be formed by the end of week 2. Make sure you obtain phone numbers and email addresses for each other. If you think finding meeting times will be difficult, you may want to start dividing the workload into subgroups.

Topics: You are free to choose your own questions. The questions may be related to your major or some other topic of interest. You should choose a topic so that it will be straightforward to gather the data. The easiest approach will be to design an experiment to compare two groups but the only rule is to make sure the topic is interesting to your group! Be creative! We will discuss some previous topics in class, and some previous project topics can be found on the course web pages. You will want to collect lots of data and then narrow in on two hypothesis pairs later. After the topics are selected, most of the work for the projects will take place outside of class.

Project Reports: The goal of the project reports is to keep you thinking about the projects as the term progresses. Keep in mind that your project may change and evolve as the course progresses. Still, with each project report I would like to hear about your progress and ideas. Turn in one project report for each team, including team members' names and *previous project reports*, preferably typed. Below are some guidelines on what I would like to see in each report.

The first project report is due Oct. 11. For this report you should identify your topic/questions of interest, the variables you plan to measure, the population you plan to draw conclusions about, the sample and sampling frame you plan to use (if applicable), and the type of study (e.g., observational study/survey or experiment).

The second project report is due Oct. 18. Your data collection techniques should be more clearly defined. If an experiment, give your tentative design. If a survey, give the preliminary questionnaire. You should indicate why this study is appropriate to answer your question and what precautions you will take (e.g., nonresponse, nonsampling bias, wording).

The third project report is due Nov. 22. You should have finished collecting your data. Include a description of your observational units, your variables, their measurement units, possible ranges/responses of these variables, as well as preliminary descriptive statistics. You should also specify two research questions/sets of hypotheses that you plan to test using your data. Indicate which set of statistical procedures you believe you will use. Include a justification for that choice of procedure. You should also outline how the remaining work will be completed (who, when).

Rough Draft (optional). If you turn in a rough draft by Nov. 29, I will review the paper, providing comments and suggestions for improving your final grade and presentation.

Final Reports: Final reports are due on or before Dec. 2. Reports must be typed. Turn in one report per group and *include previous project reports*. Incorporate computer output into the body of the paper. *Raw data should be emailed to me as an attachment*. You may assume your audience will understand all statistical terminology. Make sure the final report includes at least:

- Title page with all group member names, Fall, 2005.
- Statement of Purpose: One to two sentences outlining the topic of the study

I. Introduction

Why did you choose this topic? What did you expect to find? Have similar studies been done elsewhere? Why should the reader be interested in your results and continue reading?

II. Data Collection Methods

How did you collect the data? What were the observational units? What groups did you compare, how did you find them/form them? Type of study? What was your response variable? How were these variables measured? What additional “controls” did you exert on the study? (e.g., did you only observe people writing or did you take any behavior such as throwing a football as indication of handedness?) Any “operational definitions”? (e.g., did you field test any of the questions on a test group to see if the wording was clear?). Be especially clear on the role of randomness in your study. Are there any other potential sources of sampling or non-sampling errors? Any other unexpected results? Did anything go wrong during the course of the study? (Note: You can never give me too much detail in this section!)

III. Analysis of Results

Descriptive Statistics: You will need to make choices as to which numerical and graphical summaries are most relevant. Make sure you integrate the computer output into the body of the report and include discussions of how you are interpreting the message in these summaries. In your discussion you should fully describe your sample, sample size, and report the sample statistic and whether it supports your conjecture. Make sure all figures and graphs are clearly labeled.

Inferential Statistics: In carrying out the test(s) of significance, remember to: state your hypotheses in symbols and in words; justify your choice of procedures and comment on the validity of the methods (technical conditions); perform appropriate follow-up analyses (e.g., multiple comparisons, expected cell counts) and confidence intervals; state your conclusions in context. Pay particular attention to whether you can generalize your sample to a larger population and whether you can draw cause and effect conclusions.

IV. Conclusion

Summarize the results of your study. What did you learn? Did the data behave as you expected? Critique the methods used to collect the data. Is there anything you would do differently next time? How might this affect the conclusions of the study? What similar questions might someone chose to investigate in the future to build on your results?

Appendix Access to raw data (email), previous project reports

Grading Criteria for Final Report:

10%: Quality of written report

20%: Design of survey/experiment – was data collection adequately explained, were the appropriate data collected to answer the questions posed, was the topic original?

25%: Correctness of statistical analysis and checks of technical conditions

20%: Appropriateness of interpretations of the results of the statistical calculations and conclusions (is it a cause and effect relationship? what is a reasonable population?)

25%: Presentation – details to follow

See also J. Holcomb Projects at ARTIST Website

Including follow-up take-home examination questions

Holcomb, J. and Ruffer, R. (2000). Using a term- long project sequence in introductory statistics, *The American Statistician*, 54, 49-53.

Sample Project Rubric

Writeup (10 pts): Look at quality of writing, organization, lack of punctuation and grammatical errors, inclusion of required sections including statement of purpose. Graphs should be incorporated into the body of the report. Is the writing consistent and well constructed?

Design (20 pts): Was the initial data collection design adequate. Did students consider possible sources of bias and confounding? Did students appropriately use randomization or take a true random sample. Did students provide sufficient detail and clarity of their data collection plan? Did students define the population, sample, sampling frame, and response rate? Did students recognize and comment on any weaknesses in their original plan? Were any “operational definitions” clearly defined?

Analysis (25 pts): Were the statistical analyses correct and complete? Were the analyses consistent with the research questions (e.g., one-sided vs. two-sided)? Were the conclusions consistent with the analysis? Did students state and check all necessary technical conditions? Did students show their work clearly and use proper notation and terminology? Is all output clearly labeled and easy to read?

Interpretation (20 pts):

Did students interpret their graphical and numerical summaries and discuss whether or not these observations were supported by the inferential methods? Did they discuss why the inferential methods were appropriate for their research question? Did students make proper, well-supported conclusions? Did students provide suggestions for the results that they observed? Did students make recommendations and/or include suggestions for future studies?

Presentation (25 pts):

Did the students clearly introduce and overview the project? Was the presentation well organized, well paced, sufficiently loud, and easy to follow? Was the presentation clear and understandable? Did the visual aids enhance the talk? Were they readable (font size, amount of text per slide)? Were there enough? Did they present their final conclusions clearly? Was there sufficient eye contact? Did they talk freely instead of reading from their notes? Did they show sufficient enthusiasm for what they were presenting? Did they adhere to the time limit?

Sample Project Rubric (Emenaker, 1999) – Holistic

4 points: Exemplary Response

All of the following characteristics must be present.

- The answer is correct.
- The explanation is clear and complete.
- The explanation includes complete implementation of a mathematically correct plan.

3 points: Good Response

Exactly one of the following characteristics is present.

- a. The answer is correct due to a minor flaw in plan or an algebraic error.
- b. The explanation lacks clarity
- c. The explanation is complete

2 points: Inadequate Response

Exactly two of the characteristics in the 3-point section are present OR

One or more of the following characteristics are present.

- a. The answer is incorrect due to a major flaw in the plan.
- b. Explanation lacks clarity or is incomplete but does indicate some correct and relevant reasoning.
- c. A plan is partially implemented and no solution is provided.

1 point: Poor Response

All of the following characteristics must be present.

- The answer is incorrect.
- The explanation, if any, uses irrelevant arguments.
- No plan for solution is attempted beyond just copying data given in the problem statement.

0 points: No Response

- The student's paper is blank or contains only work that appears to have no relevance to the problem.

Sample Group Project Evaluation Form (Emert, 1999)

1. Overall, how effectively did your group work together on the project? (Poorly, Adequately, Well, Extremely Well)
2. In an effective working group, each person should be an effective participant. How well did your group meet this goal? (Poorly, Adequately, Well, Extremely Well)
3. Give one specific example of *something you learned from a group* that you probably would not have learned alone.
4. Give one specific example of *something another group member learned from you* that he or she probably would not have learned alone.
5. What is the biggest challenge to group projects? How could this challenge be overcome?

AAHE 9 Principles of Good Practice for Assessing Student Learning

1. The assessment of student learning begins with educational values.

Assessment is not an end in itself but a vehicle for educational improvement. Its effective practice, then, begins with and enacts a vision of the kinds of learning we most value for students and strive to help them achieve. Educational values should drive not only what we choose to assess but also how we do so. Where questions about educational mission and values are skipped over, assessment threatens to be an exercise in measuring what's easy, rather than a process of improving what we really care about.

2. Assessment is most effective when it reflects an understanding of learning as multidimensional, integrated, and revealed in performance over time.

Learning is a complex process. It entails not only what students know but what they can do with what they know; it involves not only knowledge and abilities but values, attitudes, and habits of mind that affect both academic success and performance beyond the classroom. Assessment should reflect these understandings by employing a diverse array of methods, including those that call for actual performance, using them over time so as to reveal change, growth, and

increasing degrees of integration. Such an approach aims for a more complete and accurate picture of learning, and therefore firmer bases for improving our students' educational experience.

3. Assessment works best when the programs it seeks to improve have clear, explicitly stated purposes.

Assessment is a goal-oriented process. It entails comparing educational performance with educational purposes and expectations -- those derived from the institution's mission, from faculty intentions in program and course design, and from knowledge of students' own goals. Where program purposes lack specificity or agreement, assessment as a process pushes a campus toward clarity about where to aim and what standards to apply; assessment also prompts attention to where and how program goals will be taught and learned. Clear, shared, implementable goals are the cornerstone for assessment that is focused and useful.

4. Assessment requires attention to outcomes but also and equally to the experiences that lead to those outcomes.

Information about outcomes is of high importance; where students "end up" matters greatly. But to improve outcomes, we need to know about student experience along the way -- about the curricula, teaching, and kind of student effort that lead to particular outcomes. Assessment can help us understand which students learn best under what conditions; with such knowledge comes the capacity to improve the whole of their learning.

5. Assessment works best when it is ongoing not episodic.

Assessment is a process whose power is cumulative. Though isolated, "one-shot" assessment can be better than none, improvement is best fostered when assessment entails a linked series of activities undertaken over time. This may mean tracking the process of individual students, or of cohorts of students; it may mean collecting the same examples of student performance or using the same instrument semester after semester. The point is to monitor progress toward intended goals in a spirit of continuous improvement. Along the way, the assessment process itself should be evaluated and refined in light of emerging insights.

6. Assessment fosters wider improvement when representatives from across the educational community are involved.

Student learning is a campus-wide responsibility, and assessment is a way of enacting that responsibility. Thus, while assessment efforts may start small, the aim over time is to involve people from across the educational community. Faculty play an especially important role, but assessment's questions can't be fully addressed without participation by student-affairs educators, librarians, administrators, and students. Assessment may also involve individuals from beyond the campus (alumni/ae, trustees, employers) whose experience can enrich the sense of appropriate aims and standards for learning. Thus understood, assessment is not a task for small groups of experts but a collaborative activity; its aim is wider, better-informed attention to student learning by all parties with a stake in its improvement.

7. Assessment makes a difference when it begins with issues of use and illuminates questions that people really care about.

Assessment recognizes the value of information in the process of improvement. But to be useful, information must be connected to issues or questions that people really care about. This implies assessment approaches that produce evidence that relevant parties will find credible, suggestive, and applicable to decisions that need to be made. It means thinking in advance about how the information will be used, and by whom. The point of assessment is not to gather data and return “results”; it is a process that starts with the questions of decision-makers, that involves them in the gathering and interpreting of data, and that informs and helps guide continuous improvement.

8. Assessment is most likely to lead to improvement when it is part of a larger set of conditions that promote change.

Assessment alone changes little. Its greatest contribution comes on campuses where the quality of teaching and learning is visibly valued and worked at. On such campuses, the push to improve educational performance is a visible and primary goal of leadership; improving the quality of undergraduate education is central to the institution’s planning, budgeting, and personnel decisions. On such campuses, information about learning outcomes is seen as an integral part of decision making, and avidly sought.

9. Through assessment, educators meet responsibilities to students and to the public.

There is a compelling public stake in education. As educators, we have a responsibility to the publics that support or depend on us to provide information about the ways in which our students meet goals and expectations. But that responsibility goes beyond the reporting of such information; our deeper obligation – to ourselves, our students, and society -- is to improve. Those to whom educators are accountable have a corresponding obligation to support such attempts at improvement.

Authors: Alexander W. Astin; Trudy W. Banta; K. Patricia Cross; Elaine El-Khawas; Peter T. Ewell; Pat Hutchings; Theodore J. Marchese; Kay M. McClenney; Marcia Mentkowski; Margaret A. Miller; E. Thomas Moran; Barbara D. Wright

This document was developed under the auspices of the AAHE Assessment Forum with support from the Fund for the Improvement of Postsecondary Education with additional support for publication and dissemination from the Exxon Education Foundation. Copies may be made without restriction.

Common Weakness in Assessment (P. Holmes, RSS Centre for Statistical Education)

- Tasks do not match stated outcomes
- Criteria do not match the tasks or outcomes
- The criteria are not known to the students
- Students do not understand the criteria
- Overuse of one mode of assessment such as written exams, essays, or discrete problems
- Overload of students and staff
- Insufficient time for students to do the assignments
- Too many assignments with the same deadline
- Insufficient time for faculty to grade the assessments
- Absence of well defined criteria so consistency is difficult to achieve
- Unduly specific criteria which create a straightjacket for students and make grading burdensome for lecturers

- Inadequate of superficial feedback provided to students
- Wide variations in marking between sections, between graders, or within graders
- Variations in assessment demands of different modules

Designing Assessments (P. Holmes, RSS Centre for Statistical Education)

1. What are the outcomes to be assessed?
2. What are the capabilities/skills (implicit or explicit) in the outcomes?
3. Is the method of assessment chosen consonant with the outcomes and skills?
4. Is the method relative efficient in terms of student time and faculty time?
5. What alternatives are there? What are their advantages and disadvantages?
6. Does the specific assessment task match the outcomes and guidelines?
7. Are the grading criteria appropriate?

Reflections on Process of Creating Exam (A. Rossman, Cal Poly)

- Assess what you value!
- Identify key concepts/skills
 - E.g., ability to interpret, critique news release
- Constantly be on look-out for good exam questions
 - Jot down ideas right after class, while grading HW
 - Base them on real data/studies as much as possible
- Make sure that students have seen exam-like questions
 - Ask similar questions in class, quizzes, HW
 - Revisit class activities, challenging HW problems
- Try to make time a non-factor
- Use multiple parts but try to minimize dependency on previous solutions
- Allow students to use books, notes!
 - That's how life is!
 - Sends clear message that thinking, reasoning, interpreting are what matter
- Provide studying advice, summary of most important ideas
- Give as much feedback as possible
- Always get feedback from colleague

Developing Assessment Plan

- Match (most important) instructional goals
 - Start by defining learning outcomes, measure their attainment
 - “best assessment derives from teachers’ questions about their own teaching”
- Multiple and varied indicators (use more than exams!)
 - Inter-related, complementary
- Well-defined, well-integrated throughout course
 - Detailed expectations when assigned
 - Part of the learning process
- Goals understood by students
 - Promote self-reflection, responsibility, trust
- Timely, consistent feedback
 - indicators for change, improve learning, feedback loop, reinforcement
 - make it “safe” for students to struggle

- Individual and group accountability
- Openness to other (justified) interpretations, reward thoughtfulness, creativity
- Not all at once, Not too much
- Collaborate
- Continual reflection, refinement, use of information
- * Assess what you value

Cautions

- Consider time requirements for students and instructor!
 - Easier to solve than to explain
 - With experience, become more efficient
- Provide sufficient guidance
 - Provide students with familiarity and clear understanding of your expectations
 - May not be used to being required to think!
 - Less comfortable writing in complete sentences

Challenges in Statistics Education

- Doing statistics versus being an informed consumer of statistics
- Statistics vs. mathematics
 - Role of context, messiness of solutions, computers handling the details of calculations, need to defend argument, evaluate based on quality of reasoning, methods, evidence used
- Have become pretty comfortable with lecture/reproduction format
 - Traditional assessment feels more objective
- Reduce focus on calculation
- Reveal intuition, statistical reasoning
- Require meaningful context
 - Purpose, statistical interest
 - Meaningful reason to calculate
 - Careful, detailed examination of data
- Use of statistical language
 - Meaningful tasks, similar to what will be asked to do “in real life”