<u>SPOTLIGHT ON RESEARCH:</u>
"SHARE YOUR RESEARCH ON TEACHING AND LEARNING STATISTICS"

Elementary Preservice Teachers' Conceptions of Variability

Daniel L. Canada, Ph.D.
Eastern Washington Univ.
Email: canada@cheerful.com
Tele: (509) 359-6074

Abstract:

My current research focuses on the following three research questions:

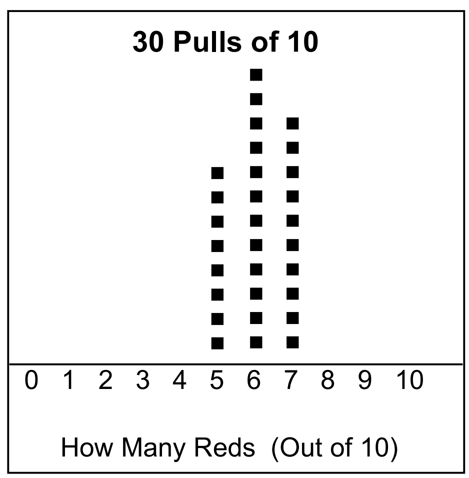
[1] What are the components of a conceptual framework that help characterize Elementary Preservice Teachers' (EPSTs') thinking about variability?
[2] How do EPSTs' conceptions of variability before an instructional intervention compare to those conceptions after the intervention?
[3] What tasks are useful for examining EPSTs' conceptions of variability in the contexts of sampling, data & graphs, and probability?

So far, collective results from survey data, interview data, and class observations have been used to describe components of an *evolving framework* useful for examining EPSTs' conceptions of variability. The three main *aspects* of the framework address how EPSTs reason in expecting, displaying, and interpreting variability. Each of the three aspects is further defined by different *dimensions*, which in turn have their own constituent *themes*. The depth in describing the evolving framework is a main contribution of this research.

What follows are three representative tasks that have been used to explore EPSTs' conceptions of variability. There is one task for each of the three contexts of sampling, data and graphs, and probability situations.

Sampling

Matt took his class to the candy container (100 Candies = 60 Red and 40 Yellow). Then he left the room. When he came back, the class claimed to have pulled 30 samples each of size 10, with replacement. They showed Matt this graph, supposedly based on their data:



Which of the following do you think is *most* likely ? Put a check mark next to it.

___ Matt's class just made up these results

___ Those are the actual results of the class samples

___ No one can have much confidence if the results are made up or not.

Explain why you think this is the most likely.

Data & Graphs

Suppose the newspaper says a movie is starting at 4:00pm. You show up at 4:00pm, but after the previews and advertisements are done, the actual movie starts at 4:20 pm! I'll call this the movie *Wait-Time*: The difference between the advertised time and the actual time the film starts.

To investigate movie wait-times, a class goes to twelve different movies at the Royal Theater and twelve different movies at the Minnow Theater. They gather the data shown (at right):

Data: (Movie Wait-Times in Minutes)

Royal Theater

7.0	7.0	7.0	11.5	10.0	11.0
7.5	7.5	13.0	14.5	13.0	11.0

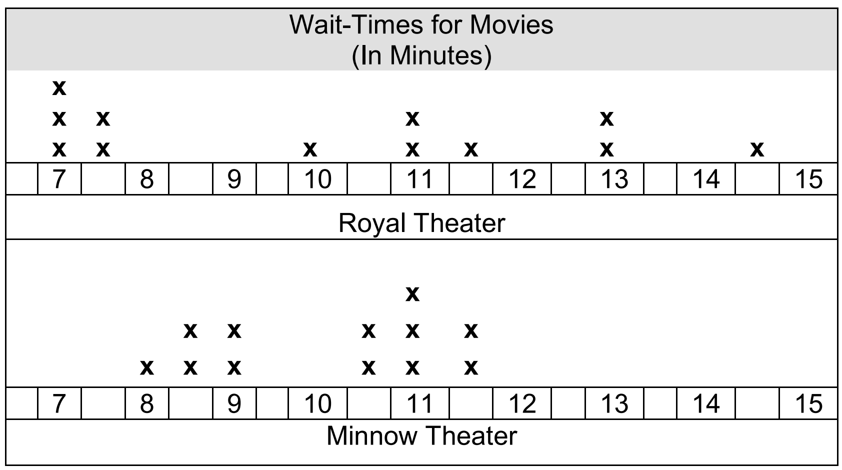
Mean = 10.0 min.
Median = 10.5 min.

Minnow Theater

8.0	8.5	8.5	9.0	9.0	10.5
10.5	11.0	11.0	11.0	11.5	11.5

Mean = 10.0 min.
Median = 10.5 min.

One student in class argues that there's really no difference in the Wait-Times of the two Theaters, since the averages are the same. Do you agree? Explain why or why not.



Probability

For homework, Mr. Blair asked each student in his class to toss a die 60 times and keep track of how many times each of the 6 sides came up. Shown are the results turned in the next day by four students.

Only one of these students actually rolled the die. The other three students just made up their results before class. What do you think is *most* likely?

Side that came up	Riki	Lynn	Lee	Pat
1	7	10	10	2
2	12	11	10	15
3	6	10	10	10
4	9	10	10	28
5	14	9	10	1
6	12	10	10	4

___i) Riki really rolled it　　___ii) Lynn really rolled it
___iii) Lee really rolled it　　___iv) Pat really rolled it
___v) No one can say. Any of the 4 students is equally likely to have really rolled it.

Explain your reasoning.

The Factors Involved in the Learning of Probability and Statistics for

Underrepresented Minority College Students

Reason for Research

- The Demand for Probability and Statistics
  - The Curriculum and Evaluation Standards for School Mathematics Recommends Increase
  - 73% of Business Schools Require Two Semesters (The American Assembly of Collegiate Schools of Business)
  - People of Color Underrepresented in Mathematical Intensive Careers
- Contributing to Society
- Understanding Published Research

The Achievement Gap

"There's tremendous implications for the future of race relations in the nation. If something isn't done, we'll end up with a bifurcated society."

- Dr. Ronald Ferguson

1998 ACT Scores

	Math	Pre-Elem Algebra	Alg/Coord Geometry	Plane Geom/Trig
African American	16.9	8.3	8.3	8.3
Native American	18.6	9.4	9.0	9.5
Caucasian	21.4	11.4	10.5	11.0
Mexican American	18.6	9.5	9.1	9.5
Asian	23.4	12.3	11.7	12.1

Factors

- Language
- Lack of Role Models
- Low Expectations
- Self-Esteem
- Self-Efficacy and Stress
- Parental Involvement

Language & Reading

"Our culture [African -Americans] emphasizes the oral transmission of information," ... "White people emphasize literacy, and their children are emerged in literacy at a younger age."

- Janice E. Hale
Education Professor
Wayne State University

Lack of Role Models

- *"Many African -American's have become disillusioned about their future job prospects and the actual value of schooling and have developed a disinterested attitude leading to depressed cognitive performance and scholastic underachievement."*

  -John Ogbu

Low Expectations

"*The devalued status and overarching presumption of inferiority that society holds regarding African -Americans permeates the walls of academia, often limiting teachers' support for and expectations of African-American students.*"

-Claude Steele

Self-esteem

- Ogbu's Factor
  - Initial Terms of the Minority Group Members' Incorporation and Response
  - Strong Racial Identity + Futility + High Levels of Ogbu Factor in African -American Males = low GPA
  - Strong Racial Identity + High Levels of Ogbu Factor + Low Futility + High Salience in African - American Females = High Achievement
- Acting "White"

Self-esteem(cont'd)

- Steele's Stereotype Threat
  - Judged Negatively for Academic Performance
  - Devalue School and Underperformance

Self-efficacy and Stress

- High Risk Neighborhoods

  "This state of condition is linked to self - efficacy " (Cunningham et al., 2002).

- Products of Their Environments

  "Contextual stress influences self -esteem which is related to academic outcomes" (Cunningham et al., 2002).

Parental Involvement

"Since the 1960s…the percentage of African-American children living with one parent has risen to 63 percent, with 92 percent of these families headed by a female."

-Battle & Scott

Inadequate Enrollment of Minorities in Mathematics

"The course taking patterns show a disproportionate number of African-American students taking lower level mathematics classes as compared to their White peers."

-J.D. House

Inadequate Enrollment of Minorities in Mathematics

- Recruitment
- Eurocentric Elitism and Social Stratification
- Teacher Expectations
- Course Taking and Curriculum
- Tracking
- Teaching Methods
- Isolation from Other Minorities

Common Difficulties in Mathematics

- Rational Numbers (Fractions, Decimals, & Percentages)
- Verbal Problem to Mathematical Models
- Fear and Anxiety (Roberts & Saxe, 1982)

Teaching Methods for Probability and Statistics

- Cognitively Guided Instruction (CGI)
- Cooperative Learning
- Question Posing
- Simulations
- Real Life Problems
- Visual
- Remove Irrelevant Information
- Exploratory Data Analysis
- Descriptive Statistics
- Common Misuses
- Strategies for Rational Numbers
- Confront Common Errors
- World Views

References

- Battle, J., & Scott, B. M. (2000). Mother-only versus father-only households: educational outcomes for African American males. *Journal of African American Men*, 5, 93-116.
- Cunningham, M., Hurley, M., Foney, D., & Hayes, D. (2002). Influence of perceived contextual stress on self-esteem and academic outcomes in African American adolescents.
- Roberts, D. M. & Saxe, J. E. (1982). Validity of a statistics attitude survey: a follow-up study. *Educational and Psychological Measurement*, 42, 907-911.

Is Assigning Homework Prior to Lecture More Effective in Statistical Learning than Assigning Homework After Lecture?

Richard A. Drapeau, Professor of Business Statistics, Lamar University,
P.O. Box 10033, Beaumont, Texas, 409-880-8653, Richard.drapeau@lamar.edu

Abstract

This session presents the results of two teaching methods for introductory statistics - in one class students submitted homework prior to lecture and in the other class students submitted homework after lecture. It is hypothesized students assigned homework before lecture demonstrate higher statistical learning than students assigned homework after lecture.

Motivation for Study

Business students at Lamar University are required to take two semesters of statistics which are offered at the junior level. The first course covers descriptive statistics, basic probability, discrete and continuous probability distributions, sampling theory, estimation, and introduction to test of hypothesis (single population mean and single population proportion). The second course covers: a review of estimation and tests of hypothesis, expands hypothesis testing to include single population standard deviation, two population means, two population proportions, two population variances, ANOVA models, nonparametric models for two or three populations of quantitative data, chisquared models for categorical data, simple/multiple linear correlation/regression analyses, and time series.

In Fall 2000, I taught three sections of the first statistics course. The average on the first examination was 53 with 58 percent of the students failing the examination. Because students were provided three sets of multiple choice examinations as a study guide, I refused to curve grades. Immediately, students started dropping the course. By the end of the semester, 78 percent of the students either dropped or failed the course. The department chair called me to his office and asked me to "lower the bar without watering down the course." According to the chair, students' major complaints were the difficulty of the midterm examinations.

The Study

For Spring 2001, I redesigned the course. Instead of administering quizzes between the three major midterm examinations and requiring a comprehensive final examination, the course was redesigned by grading homework, administering MORE quizzes, eliminating all midterm examinations, and requiring a comprehensive final examination.

I felt I was both "watering down the course" and "lowering the barrier" because I was lowering performance standards to accommodate students at the expense of learning. I decided I needed to do something in the course to give the lower expectations meaning. I decided that one of the two sections of the first semester course would work homework before my lecture. The other section would work homework after my lecture. The syllabus stated:

> The syllabus contains the reading assignment for each class period. Students in both classes are expected to read the assigned pages BEFORE attending class. During this semester there will be two approaches of assigning homework. For Section 1 (10:10 am), the problems appearing with the assigned reading are to be worked BEFORE attending the class lecture. For Section 2 (11:00 am), the problems appearing with the assigned reading are to be worked AFTER attending the class lecture (however, students in this section are expected to have read the assigned reading). It is hypothesized that students in Section 1 will perform higher on the average on quizzes and the final examination than students in Section 2. During the semester, Professor Drapeau will collect ten homework problems. Students in Section 1 need to bring homework to class on Wednesday and Friday and one Monday (March 5). Students in Section 2 need to bring homework to class of Friday and Monday and one Wednesday (March 7). Professor Drapeau will not accept late homework. Each homework problem selected will be graded on a base of five points. Although Professor Drapeau will be more lenient in

grading homework from students in Section 1 than from students in Section 2, students in both sections must clearly show their work. Students who merely copy the solutions in the back of the textbook will not receive full credit. To maximize homework points, write the formula, substitute numbers, and show the final calculation. For problems in Chapters 7-10, draw pictures which clearly identify the appropriate probability. To Professor Drapeau, it is more important for students to show the process of solving a problem than to generate the final answer. If more than ten homework problems are collect, the best ten homework problems will contribute fifty points to your semester grade.

In addition to ten homework assignments, the syllabus scheduled twelve quizzes. The syllabus stated:
Twelve quizzes will be administered during the session. These quizzes are scheduled for Monday classes. The scheduled dates of these quizzes appear on the syllabus. Each quiz will consist of true/false questions, short answers, simple problems, and/or computer printouts. When the quiz is simple problems, you must show your work to obtain full credit. If you only provide the answer -- only part of the solution -- you will receive only partial credit. Each quiz will be worth 20 points. The best ten of the 12 quizzes will be used for your semester quiz score. The quizzes contribute 200 points to your final course grade. On the day of the quiz, the first 25 minutes of class time will be used to answer questions over material covered during the previous week. The second 25 minutes of class time will be used to work the quiz. Calculators are permitted on quizzes.

I decided to maintain the comprehensive final examination. The syllabus stated:
A comprehensive final examination worth 150 points will be administered on the date as specified by the University. Necessary formulae will be provided by the student bringing a HELP SHEET to the examination. For the final examination, the HELP SHEET will consist of both sides of TWO 8.5x11 inch sheet of paper. You may write anything on the HELP SHEET, but everything on the HELP SHEET is to be HAND WRITTEN. ANY ATTEMPT TO USE MORE THAN TWO, TWO-SIDED HELP SHEET WILL BE CONSIDERED CHEATING. **A STUDENT REFERRING TO THE HELP SHEET AS A "CHEAT" SHEET WILL LOSE THE PRIVILEGE OF USING A HELP SHEET.** IF YOU FORGET YOUR CALCULATOR OR HELP SHEET, YOU WILL COMPLETE THE EXAMINATION WITHOUT THE RESOURCE; YOU MAY NOT USE YOUR NEIGHBOR'S CALCULATOR OR HELP SHEET.

A comparison of the course structure for Fall 2002 and Spring 2001 appears in the Table 1.

Table 1
Comparison of Traditional vs. Modified First Semester Statistics Course

Activity	**Fall 2000**	**Spring 2001**
Homework collected/graded	None	10 assignments (collected 11) 50 points
Quizzes	Best 7 of 8 100 points	Best 10 of 12 (actually 10 of 14) 200 points
Midterm Exams	3 300 points	None
Comprehensive Final	1 200 points	1 150 points
Total Points	600 points	400 points

The Research Hypothesis

Intuitively, students in Section 1 (pre-lecture homework) would be expected to perform better than students in Section 2 (post-lecture homework) because they had to carefully read the text and work assigned homework before I lectured on the material. I would expect that this "pre-preparation" would (1) develop critical thinking skills, (2) facilitate comprehension of statistical concepts and (3) provide learning to perform better on quizzes and the comprehensive final examination. This reasoning suggests a one-tail test: H_0: $\mu_1 \leq \mu_2$ and H_0: $\mu_1 > \mu_2$ where means are compared for homework, quizzes, and each component of the comprehensive final examination.

Statistical analysis used Oneway ANOVA and the General Linear Models of SPSS. Both these models test the null hypothesis that all means are equal with the alternative that at least one mean is different. With only two samples, the direction of difference can be determined by comparing the two sample means. It must be noted that these statistical models are based upon probability samples. Students in these two classes were actually convenience samples. The assumption, regardless of how weak it may be, is that these convenience samples are as representative of the two populations as are random samples. A level of significance of 0.05 was selected.

Findings

Table 2 presents the analyses of five Oneway ANOVA analyses comparing the difference between the two sections in mean performances on homework, quizzes, and each component of the comprehensive final examination.

Table 2
Oneway ANOVA
Mean Performances on Assessment Instruments
BUAL 3310, Section 1 and Section 2

Variable	Homework (11)	Quizzes (14)	Final Exam Part I (True/False)	Final Exam Part II (Problems)	Final Exam Both Parts
Maximum Points	55	280	50	100	150
Section 1 (n_1=15)	32.9	175.5	29.4	44.3	73.7
Section 2 (n_2 = 14)	22.5	168.4	31.4	45.2	76.6
P-value	0.016	0.632	0.336	0.881	0.694

Section 1 (pre-lecture homework) generated a higher average on homework and quizzes but a lower average on both parts of the comprehensive final examination than did Section 2 (post-lecture homework). If these two small samples are treated as probability samples, then the only difference which is statistically significant at the 0.05 level of significance is the mean homework grades between the two sections. As noted in the syllabus (and as practiced in class during the semester), I graded homework from Section 1 more leniently than homework from Section 2. Therefore, the statistical difference could be easily attributable to the way I graded homework from the two sections.

Although there is no statistically significant difference between mean performances on each of the two parts of the final, the performance on the final examination was disastrous! Part I of the comprehensive final examination consisted of 25 true/false questions testing theory. If students marked a question false, they had to correct the statement to make it true. Out of 50 possible points, the mean grade was 29.4 points (58.8%) in Section 1 and 31.4 points (62.8%) in Section 2. Part II of the comprehensive final examination

consisted of problems worked longhand for which partial credit was assigned. Performance on Part II was worse than on Part I; out of 100 possible points, the mean grade was 44.5 points (44.5%) in Section 1 and 45.2

point (45.2%) in Section 2. For the combined parts, the mean grade was 73.7 points (49.1%) and 76.6 points (51.1%) in Section 2. It was obvious that the weekly quizzes only encouraged short-term learning, so students were not prepared to take a comprehensive final examination. Because performance on the comprehensive final examination was so poor, the higher number between semester percentage prior to the comprehensive final examination and semester percentage after the comprehensive final examination was used to assign semester grades. For all but one student, performances on the comprehensive final examination reduced semester percentages, so course grades were based upon performances entering the comprehensive final examination. The one student who improved his semester percentage was failing before and after the comprehensive final examination.

Realizing that performance could be influenced by aptitude, GPA was used as a measure of aptitude and a Oneway ANOVA was performed on GPA of students in both sections. This analysis is presented in Table 3.

Table 3
Mean GPA of Students in BUAL 3310, Section 1 and Section 2

	Mean GPA
Section 1 (n_1 = 15)	2.74
Section 2 (n_2 = 14)	2.94
P-value	0.384

The mean GPA of students in Section 1 is lower than the mean GPA of students in Section 2. Assuming the samples to be probability samples, the difference is not statistically significantly at the 0.05 level of significance.

Realizing that performance could be influenced by the amount of time students work on jobs outside of school, a Oneway ANOVA was performed on the hours worked per week (HWPW) of students in both sections. This analysis is presented in Table 4.

Table 4
Mean Hours Worked Per Week for Students in BUAL 3310, Section 1 and Section 2

	Mean HWPW
Section 1 (n_1 = 14)	35.6
Section 2 (n_2 = 14)	26.9
P-value	0.153

The mean HWPW of students in Section 1 is higher than the mean HWPW of students in Section 2. Assuming the samples to be probability samples, the difference is not statistically significantly at the 0.05 level of significance.

Five linear models analyzing the difference in means of each of the five assessment instruments were run using GPA as a covariate; these are referred to as Model 2. Five linear models analyzing the difference in means of each of the five assessment instruments were run using HWPW as a covariate; these are referred to as Model 3. Finally, five linear models analyzing the difference in means of each of the five assessment instruments were run using both GPA and HWPW as covariates; these are referred to as Model 4. Table 5 presents the results of these analyses.

Table 5
Mean Performances on Assessment Instruments Section 1 and Section 2
Model 1 – Compare Only Performance Model 2- Performance Adjusted for GPA
Model 3 – Performance Adjusted for Hours Worked Per Week
Model 4 – Performance Adjusted for GPA and Hours Worked Per Week

Variable	Homework (11)	Quizzes (14)	Final Exam Part I (True/False)	Final Exam Part II (Problems)	Final Exam Both Parts
Section 1[1]	32.9	175.5	29.4	44.3	73.7
Section 2[2]	22.5	168.4	31.4	45.2	76.6
Model 1 P-value	0.016	0.632	0.336	0.881	0.694
Model 2 P-value	0.001	0.115	0.452	0.561	0.825
Model 3 P-value	0.026	0.619	0.569	0.600	0.768
Model 4 P-value	0.001	0.001	0.738	0.112	0.220

[1] **n_1 = 15 for Models 1 and 2, n_1 = 14 for Models 3 and 4**
[2] **n_2 = 14 all models**

As with the Oneway ANOVA, the General Linear Models consistently found the difference in mean performances on homework between the two sections to be significantly different. Using a five percent level of significance, mean homework of Section 1 was higher than mean homework of Section 2. However, as previously explained, grading of Section 1 homework was more lenient than grading of Section 2 homework. Model 4 – comparison of means adjusted for both GPA and HWPW – indicated mean quiz grade of Section 1 to be significantly higher than the mean quiz grade of Section 2.

As noted in the discussion of the Table 2 analysis, mean performance on each part of the comprehensive final examination in Section 1 was lower than in Section 2; however, at the five percent level of significance this difference was not statistically significant. As previously noted, students were not prepared to take a comprehensive final examination.

<u>Summary and Conclusions</u>

This study was motivated by the need to "lower the bar" without "watering down the course." Students in one section of the first junior level business statistics course were required to work homework before class lecture while students in another section of the same course were required to work homework after class lecture. Assessment measures included (1) 11 homework assignments, (2) 14 quizzes, (3) a comprehensive final examination consisting of 50 points of true/false questions and 100 points of problems worked long-hand. Although the two sections represent convenience samples, they were "assumed" to represent all students who take the first statistics class. Therefore, statistical analysis was performed using Oneway ANOVA and General Linear Models; level of significance was set at 0.05. Four models were tested: (1) simple comparison of means on the five performance measures, (2) comparison of means adjusted for GPA, (3) comparison of means adjusted for HWPW, and (4) comparison of means adjusted for both GPA and HWPW.

Homework was the only performance measure that was statistically significant on all four models. Students in Section 1 preformed statistically higher, on the average, than students in Section 2. As previously noted, grading of homework in Section 1 was more lenient than grading of homework in Section 2.

Quizzes were statistically significant for the model which compared mean performance adjusted for GPA and HWPW. Mean performance in Section 1 was higher than mean performance in Section 2.

Performance on each of the two parts of the comprehensive final examination was lower by students in Section 1 than by students in Section 2. However, the difference was not statistically significant for any of the four models. As previously noted, weekly quizzes only focused on short-term learning and did not prepare students for a comprehensive final examination. The average grade on the comprehensive final examinations in both sections was around 50%. It is apparent that "lowering the barrier" was achieved but "without watering down the course" was not achieved.

Limitations of the study include (1) use of statistical models for nonprobability samples and (2) small sample sizes. Additional analysis will include (1) impact of the number of courses students took and (2) impact of number of college math courses completed before taking the first business statistics course. Including these additional explanatory variables may identify differences in performances between the two sections.

Summary of *P*-value Survey Research

Sharon J. Lane-Getaz, University of Minnesota
lane0139@umn.edu

Reviewing the literature across the areas of education, psychology, statistics, and statistical and mathematics education illuminates the difficulties people have with the logic of inference. Within this literature the common indicator of statistical significance, the P-value, is fraught with controversy and confusion. Thirteen P-value misconceptions have been documented in empirical studies. In fall 2004 students in first and second courses in statistics responded to survey items developed based on this review of the research literature about P-values. The survey development and validation is discussed, along with summaries of students' correct reasoning or misconceptions about P-values based on this preliminary pilot.

1. Literature Review

Misinterpretations of *P*-values appear to be common among statistics students and some experienced researchers as well. A compilation of *P*-value misconceptions from these six empirical studies is listed in Table 1. This spotlight session highlights results of a pilot of a seventeen item survey designed to diagnose students' conceptions and misconceptions about *P*-values based on the research literature.

Table 1: Compilation of P-value misconceptions documented in empirical research

P-value misunderstandings and misinterpretations
1. *Significance level confusion:* The predetermined significance level, α and *P*-value are confused (Vallecillos, et al., 1992; Williams, 1997, 1998, 1999).
2. *P-value always low:* The belief that all *P*-values are small in value (Williams, 1998, 1999).
3. *P-value versus test statistic confusion:* The student confuses or doesn't understand the relationship between *P*-values and test statistics (Williams, 1998, 1999).
4. *Confusion of the converse (Causality):* Interprets the *P*-value—which is $P(D\|H_o)$—as $P(H_o\|D)$; Interprets small *P*-values as probability *the null hypothesis is true* (Brewer, 1985; Falk, 1988; Haller & Kraus, 2002; Pollatsek, et al., 1987; Vallecillos & Holmes, 1994).
5. *Significance testing confusion*: The logic or language of significance testing presents an obstacle to understanding and interpreting *P*-values (Batanero, 2000; Vallecillos, et al., 1992).
6. *Sample size dependence or effects:* Failure to recognize *P*-values are dependent on sample size; related to power of the test or the treatment effect (Mittag & Thompson, 2000; Wilkerson & Olsen, 1997).
7. *Confusion between sample effects and population effects:* Reflects a belief that the *P*-value is the probability of sample effects, rather than population effect under the null (Mittag & Thompson, 2000)
8. *Odds-against-chance fantasy:* Reflects a belief that significant *P*-values can be used to decide to accept or reject the idea that *chance caused* the experimental results obtained (Carver, 1978).
9. *Illusion of probabilistic proof (Determinism):* Reflects belief that small *P*-values ($p < \alpha$) justify a definitive statement; i.e., *outcome approach* (Oakes, 1986; Falk, 1988; Konold, 1989; Cohen, 1990).
10. *Valid hypothesis fantasy:* Reflects a belief that the *P*-value is the probability that the research hypothesis; i.e., *alternative hypothesis is true* (Carver, 1978; Brewer, 1985; Oakes, 1986; Haller & Kraus, 2002; Vallecillos & Holmes, 1994).
11. *Reliability / Replicability fantasy:* Reflects a belief that the *P*-value is related to reliability or that the repeatability of the research results is 1 – *P*-value (Carver, 1978; Oakes, 1986; Haller & Krauss, 2002; Mittag & Thompson, 2000).
12. *Probability H_a is "wrong:"* Reflects a belief that the *P*-value is the probability that the research hypothesis is "wrong" (Oakes, 1986; Vallecillos & Holmes, 1994; Williams, 1998, 1999; Brewer, 1985).
13. *P-value and Type I error:* Failure to differentiate between the *P*-value and Type I error rates (Garfield & Ahlgren, 1988; Haller & Kraus, 2002; Mittag & Thompson, 2000).

2. Survey and Analysis

• *How do beginning and more experienced students of statistics differ in their conceptions and misconceptions about P-values?*

Tables 2 through 5 detail the number and percentage of correct responses to each of the 17 items by course. The tables mirror the four sections of the *P*-value survey: Defining *P*-values, Using *P*-values, Interpreting *P*-values and Drawing Conclusions from *P*-values. The *P*-value conception or misconception being assessed is described in the left column of these tables. In the right columns are the number of correct answers for the item by course. Each of the four survey sections has a problem context. The scenario for each section precedes the results.

Table 2: Defining P-values—results by course

Scenario 1: A research article gives a P-value of .001 in the analysis section. Do you think the following definition is true or false?

	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
1. Null hypothesis is true (False)	41 39%	28 43%	68 63%	29 51%
2. Formal definition (True)	56 54%	44 68%	85 79%	43 75%
3. Simulation definition (True)	52 16%	33 51%	62 57%	35 61%
4. Lay (informal) definition (True)	56 54%	45 69%	57 53%	37 65%
5. Population Proportion (False)	41 40%	31 48%	86 80%	46 81%

See note below.

Table 3: Using P-values—results by course

Scenario 2: District administrators of an experimental program similar to Head Start are interested in knowing if the program had an impact on reading readiness of first graders. Assume that the historical, pre-implementation mean Reading Readiness score for all first graders is 100 and the population standard deviation is 15. A random sample of current first graders who have been through program scored a mean Reading Readiness of 102.

	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
6. Sample size impact (Valid)	63 61%	40 62%	60 56%	35 61%
7. Results due to chance (Invalid)	37 36%	12 19%	24 22%	19 33%
8. "Odds" against chance (Invalid)	25 24%	13 20%	14 13%	6 11%
9. Stochastics definition (Valid)	69 67%	47 72%	82 76%	46 81%

See note below.

Table 4: Interpreting P-values—results by course

Scenario 3: An ethical researcher is hoping to show that his new hair growth treatment had statistically significant results. How should this researcher interpret results from the research study?				
	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
10. Rareness measure (Valid)	53 52%	36 55%	76 70%	41 72%
11. Test statistics confusion (Invalid)	40 39%	24 37%	27 25%	11 19%
12. Converse is true (Invalid)	34 33%	39 60%	57 53%	33 58%
13. Large P-value significant (Invalid)	44 43%	34 52%	54 50%	20 35%

Note: Lower = lower division undergraduate statistics course; Upper = upper division undergraduate statistics course; 1st Masters = masters level first course in statistics; 2nd Doctoral = doctoral level second course in statistics

Table 5: Drawing Conclusions from P-values—results by course

Scenario 4: A researcher conducts an appropriate hypothesis test where she compares the scores of a random sample of students' SAT scores to a national average (500). She hopes to show the students' mean score will be higher than average. The researcher finds a P-value for her sample of .03.				
	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
14. Reliability (Invalid)	30 29%	24 37%	60 56%	27 47%
15. Valid Hypothesis (Invalid)	48 49%	40 62%	83 77%	41 72%
16. Wrong (Invalid)	40 39%	36 55%	67 62%	33 58%
17. Type-I (Valid)	57 55%	46 71%	61 57%	37 65%

Note: Lower = lower division undergraduate statistics course; Upper = upper division undergraduate statistics course; 1st Masters = masters level first course in statistics; 2nd Doctoral = doctoral level second course in statistics

- *Do beginning and more experienced students of statistics share the same conceptions and misconceptions about P-values?*

The *P*-value survey data collected in the fall 2004 sheds some light on this question. Figures 1 and 2 depict the means and box plots of total correct scores cross tabulated by course. There do appear to be some differences in these samples.

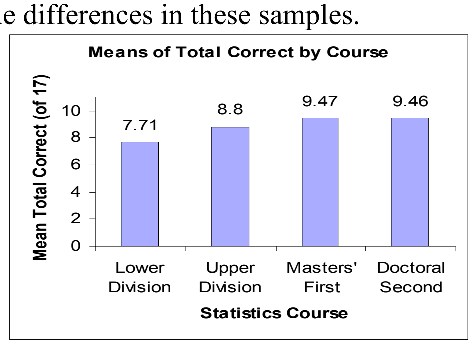


Figure 1: Mean of total items correct by course

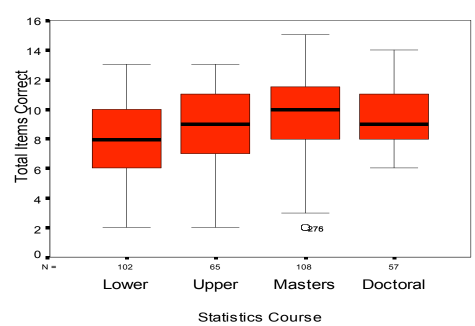


Figure 2: A comparison of medians and boxplots

Graduate students tended to get more items correct than undergraduates. There was less variation between scores among doctoral students taking their second course in statistics. In addition, there was evidence of a statistically significant difference in total scores between courses as well ($F_{.05}(3, 328) = 10.7$, $p < .001$).

3. Conclusions and implications for teaching and research

The results of this investigation indicate that this survey measures two levels of thinking: lower level statistical literacy and a higher level statistical reasoning and thinking. *P*-value misconceptions seem to require a deeper level of processing about *P*-values. Graduate respondents tended to answer correctly to more of the higher order thinking items.

There were four items that functioned counter-intuitively and require additional item development efforts. More qualitative detail is needed to shed light on why these items are discriminating so poorly. Either these items need to be improved or eliminated from the survey altogether. Future item development and modification should include cognitive interviews in which respondents "talk aloud" as they conduct the survey. These sessions can be videotaped to capture all of their reactions without intervention. This technique was used during the initial development process as well. In addition, some respondents should be interviewed after they have taken the survey to better understand how the survey is received. Some qualitative data and analysis may add the needed depth of information to further develop the instrument.

In addition to serving as a research tool, this survey can fulfill a practical role as a diagnostic tool for classroom use. For example, in spring 2005 this survey was used as a formative assessment of students' understanding of *P*-values after completion of an inference unit. The results helped the instructor determine how to target a final review on *P*-values, prior to giving students the final summative assessment. The survey identifies subtle aspects of the *P*-value that may remain elusive after instruction. By taking the survey students' misconceptions and misunderstandings are made explicit. Once they are aware of these misconceptions, the instructor has a teaching opportunity to confront and potentially overturn students' misunderstandings.

References

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*(1&2), 75-97.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, (10)*3, 252-268.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4), 287-292.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons, Inc.

Falk, R. (1988). Conditional probabilities: Insights and difficulties. In R. Davidson, & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 292). Victoria, BC: University of Victoria.

Falk, R., & Greenbaum, C. (1995). Significance tests die hard. *Theory & Psychology, 5*(1), 75-98.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44-63.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.

Lane-Getaz, S. J. (2005). Reasoning about *P*-values to develop the logic of inference: A review of related literature. Unpublished manuscript.

Lane-Getaz, S. J. (2005). Reasoning about *P*-values: Results of a research-based survey. Unpublished manuscript.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4),14-20.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences,* Chichester: Wiley.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40*, 255-269.

Vallecillos, J. A., & Holmes, P. (1994). Students' understanding of the logic of hypothesis testing. In *Proceedings of the Fourth International Conference on Teaching Statistics.* Marrakech, Morocco: International Statistical Institute.

Wilkerson, M., & Olson, J. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology, 131*(6), 627-631.

Williams, A. M. (1999). Novice students' conceptual knowledge of statistical hypothesis testing. In J. M. Truran, & K. M. Truran (Eds.), *Making the Difference: Proceedings of the Twenty-second Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 554-560). Adelaide, South Australia: MERGA.

Williams, A. M. (1998). Students' understanding of the significance level concept. In L. Pereira-Mendoza (Ed.), *Statistical Education - Expanding the Network: Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 744-749). Voorburg, The Netherlands: International Statistical Institute.

Handout for the Spotlight Presentation in Statistics Education Research Area

An Investigation of Students' Knowledge Retention of Statistical Concepts Using Problem-Posing Methodology

Carl Lee
Central Michigan University
Mt. Pleasant, MI 48859
EMAIL: Carl.lee@cmich.edu

This presentation proposes to investigate how well students retain the knowledge of statistical concepts one semester after taking the introductory statistics using a problem-posing and interview methodology.

Students who took their introductory statistics course in the Fall of 2004 were the target population of our study. It is often to hear comments from colleagues who teach higher level courses that require introductory statistics as their pre-requisite:

" My students just completed Introductory Statistics last semester, but they do not remember any thing about statistics. They don't even have any clue about very basic statistics such as median, standard deviation, confidence interval".

We design a study to investigate, if any, what statistical concepts students retain and why they can/can't retain basic statistical concepts.

From the target population about 400 students taught by seven instructors, who all have been teaching introductory statistics for many years at Central Michigan University. Students are mostly majoring in College of Business Administration. About 45 percents are mail and 55 percent are female. Their grade levels range from freshman to senior with majority being junior.

The subjects are selected based on the following design. Six students are selected from each instructor's class. Two students are randomly selected from the A and A- level students, two from B level and 2 from C or lower. A total of 42 students are selected and are given a set of five different problem scenarios. The sampling design is a stratified random sampling design. This random sample represents students from different instructors and from different grades. It allows us to make some comparison between the level of problems they pose among different levels of grades (from A to C or lower). Students are then asked to pose any question that is related to the statistical topics covered by the problem scenario. Each student is given the same instruction before posing questions:

"Now, thinking that you are a professor, and you are going to prepare questions to test your students. Here are five different problem scenarios. Each problem scenario covers a variety of statistical concepts and topics you learned from last semester. You may pose as many questions as you can and pose some simple and some difficult and some challenging questions to test your students. You have a total about 25 to 30 minutes to pose your questions. Once you complete your problem posing for one scenario, move on to next".

Most students completed their problem posing within 25 minutes. An interview is then conducted with each student to discuss the type of questions they pose, the purpose of posing each question, and what is the correct answer for each question they posed.

The interview lasted for about 30 to 40 minutes. The interviews were completed recently. We begin to edit the video clips and to record their posed questions. The following are some examples students posed for each problem scenario.

(A) Probability/Conditional Probability
The following is a table that gives the preference of 1000 students for buying brand name shoes:

	Brand Name				
Gender	Nike-Sports	Nike-Nonsport	Adidas-Sports	Adidas-Nonsports	Total
Female	150	180	150	50	500
Male	210	60	100	100	500
Total	360	240	250	150	1000

One student is chosen at random from this group of 1000 students. Pose as many questions as you can to test different concepts related to probability and conditional probability.

Here is an example of questions posed by a "D" student and an "A" student:

A 'D' student posed the following questions:

(1) What brand to you wear?
(2) How does your choice relate to overall?
(3) Is your brand prefered over gender?
(4) How does your brand compare to opposite gender same brand?
(5) Does the gender affect brand or vice versa?

An 'A' student posed the following questions:
(1) Out of the 1000 students, what is the probability that a student is wearing Adidas sports shoes?
(2) What is the probability that a male student is wearing Nike-Nonsport shoes?
(3) What is the probability of finding a female wearing Adidas-nonsport shoes?
(4) What is the probability that if someone is wearing Adidas-sports shoes then the person is male?
(5) What is the probability that if someone is male then he is wearing Adidas-nonsports shoes?

The "D" student did not seem to have any idea about any statistical concepts related to this problem. The student basically posed very descriptive questions about opinions. No question is related to probability concepts.

The "A" students, although was able to pose questions about probability of an event (Q1) , conditional probability (Q4 and Q5), the questions (2) and (3) are confusing. For example, Q2 can be interpreted as probability of intersection:

" probability that the student is a male and wearing Nike-Nonsport"

or a conditional probability

"If the student is a male, find the probability this male wearing Nike-Nonsport shoses".

It is noticed that similar language has been used by some other students as well.
When asking them what statistical concepts they wanted to test when they first read the problem scenario, no student were able to describe in terms of the statistical terminology such as "testing students if they can find the probability of intersection or union of two events".

When carefully examining the questions posed, it becomes clear that students could only pose procedural questions or one-step solution questions. We rarely notice that students could pose questions that are conceptual or require a good understanding of more than one concepts. No student posed any questions related to, for example, mutually exclusive and sample space. Only a few could pose questions related to intersection, union or conditional probability. Even so, many of such questions were confusing similar to the "A" student described above.

We are beginning to analyze the questions and try to make sense out of them. The following is the five scenarios used in this study.

Problem Posing Interview Protocol

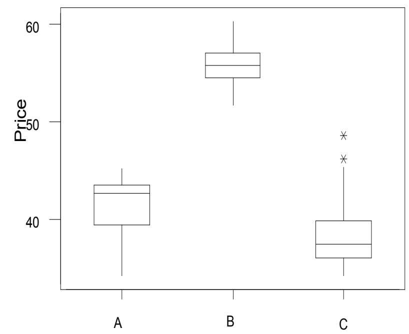
Imagine you are a statistics instructor. You need to prepare for test questions to test if students have learned the knowledge and skills you want them to learn and can challenge the best students in your class.

(A) Probability/Conditional Probability
The following is a table that gives the preference of 1000 students for buying brand name shoes:

	Brand Name				
Gender	Nike-Sports	Nike-Nonsport	Adidas-Sports	Adidas-Nonsports	Total
Female	150	180	150	50	500
Male	210	60	100	100	500
Total	360	240	250	150	1000

One student is chosen at random from this group of 1000 students. Pose as many questions as you can to test different concepts related to probability and conditional probability.

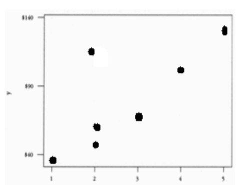
(B) The following summaries are from 40 days of three stocks. In the box plots, they are labeled as A, B and C. In the Descriptive Summary, they are labeled as X, Y and Z. You may pose questions related descriptive statistics, estimation as well as hypothesis testing.
(Note: A,B,C do not necessarily match the order of X,Y,Z).

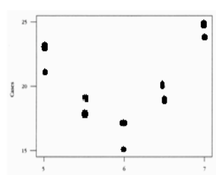


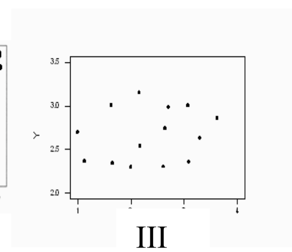
Variable	N	Mean	Median	StDev	Q1	Q3
Stock X	40	41.7	______	2.86	39.5	43.6
Stock Y	40	55.9	54.8	1.89	54.6	57.2
Stock Z	40	38.4	______	______	36.0	39.9

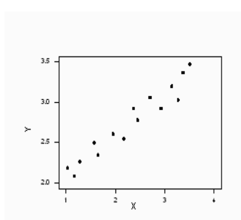
(C) The saving accounts in a large city are generally not normally distributed, instead, they are very skewed to the right, that is many accounts have small savings except a few. Suppose the saving accounts for this city have the average $\mu = \$3000$ and s.d. $\sigma = \$3000$. Bank A randomly takes 5 accounts daily and computes the average of the five for 500 days. Bank B randomly takes 50 accounts daily and computes the average of the 50 for 500 days. Use this information to pose questions related to distribution and sampling distribution of sample mean.

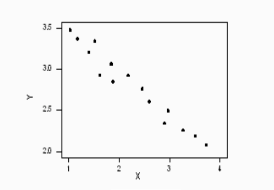
(D) On the right are six patterns of relationships between two variables. Use these plots to pose questions

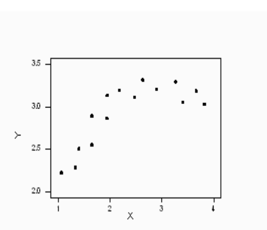

I


II


III


IV


V


VI

(E) In a statistics class, students were asked to conduct a project to study if arm length can predict the height. 40 students were randomly selected and their arm lengths and heights were recorded.

The computer output is given below.

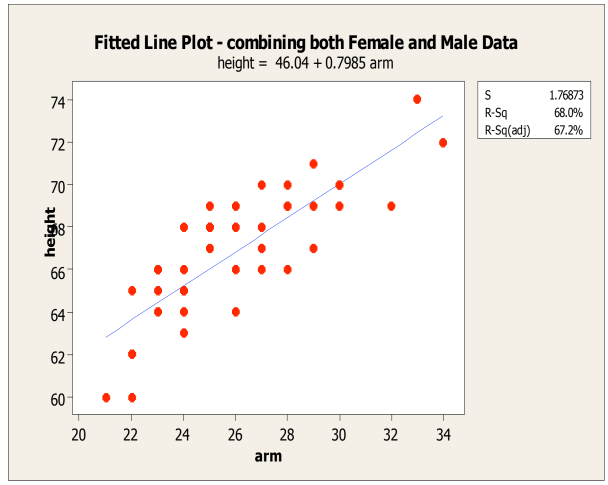
Regression Analysis: height versus arm length

```
The regression equation is:

       height = 46.04 + 0.7985 arm

Predictor      Coef    SE Coef       T       P
Constant     46.040     2.338    19.69   0.000
arm         0.79849   0.08884   ______  ______

S = 1.76873   R-Sq = 68.0%   R-Sq(adj) = 67.2%
```



An Attempt to Move From "Sadistics" to Statistics

Nyaradzo Mvududu, Ed.D.
Seattle Pacific University

Why look at this?

- Students "fear" of statistics
- Student concerns about being successful in a statistics course
- Student failure to see the relevance of statistics in their lives

My focus in the classroom

- What non-cognitive factors impact student success?
  - Anxiety
  - Attitude
  - Student perception of the classroom environment
  - Hope

Measures used at the beginning and at the end of the quarter

- The Statistics Anxiety Scale (SAS; Pretorius & Norman, 1992)
  - 5 point Likert scale
  - higher scores indicate higher anxiety
  - added items that look specifically at test anxiety

- Attitude Towards Statistics survey (ATS; Wise, 1985)
  - 5 point Likert scale
  - higher score indicate more positive attitude
  - 2 subscales – Field and Course

- The Constructivist Learning Environment Survey (CLES; Taylor, Fraser, & Fisher, 1996)
  - 5 point Likert scale
  - 2 forms – actual and preferred
  - 5 subscales – Personal relevance, Critical voice, Uncertainty, Shared control & Student negotiation

- The Hope Scale (Snyder, Harris et al., 1991)
  - 4 point Likert scale
  - higher score indicate more hope
  - 2 subscales – Agency and Pathways

Findings

Relationships

	Field	Course	Total	Critical voice	Student negotiation
General Statistics anxiety	.64 (.52)	.91 (.84)	.81 (.75)	(-.57)	(-.50)
Test anxiety		.59		(-.55)	(-.50)
# of previous stats classes	.27	.29	.31		
Personal relevance (perception)	.42		.38		

Values in ( ) are for a sample of graduate students only

For graduate students: significant correlation between perceptions of learning environment (uncertainty) and Test Anxiety at the end of the course ($r = .48$)

Factors related to Hope

	Agency	Pathways
Course		.42 (.47)
Total attitude		.41
Uncertainty (fit)	.47	
Student negotiation (fit)		.33 (.50)

Factors related to Statistics Performance

	Performance
Course (one tail)	.20
Critical voice (perception)	.29
Student negotiation (perception)	.30
# of previous math classes	-.30
Test anxiety (initial)	-.43
Test anxiety (end)	-.32

Change over the quarter (non of the changes statistically significant)

1. Reduction in level of general statistics anxiety – 29.93 v. 29.24- as well as test anxiety – 12.22 v. 12.00
2. Improvement in attitude towards the course (26.95 v. 27.88)
3. Minimal change in attitude towards the field of statistics (75.10 v. 74.76)
4. Improvement in overall attitude towards statistics (102.05 v. 102.63).
5. More hopeful with regards to both agency (13.68 v. 13.71) and pathways (12.90 v. 13.10).

Difference between graduate an undergraduate students

1. Graduate students had a more positive attitude towards the statistics course (30.88 v. 25.75; $\eta^2 = .12$) and overall attitude (108.88 v. 98.21; $\eta^2 = .11$).
2. Graduate students felt a closer person-environment fit with regards to Personal Relevance ($\eta^2 = .09$).
3. Undergraduate students perceived their learning environment as providing opportunities for Critical Voice (21.81 v. 21.11; $\eta^2 = .11$) and Shared Control (14.12 v 12.32; $\eta^2 = .09$).

Now what?

- Change possible
- To reduce anxiety
    - Create an environment where students are "safe" to question and to explain/justify their ideas. Also make the content relevant to the students' everyday lives
        i. Use data from "real" projects with which they are involved
        ii. More doing of statistics than learning about it
        iii. As far as possible link to other courses
        iv. Frequently check in and ask for feedback
        v. Collaboration even on tests
- May lead to higher hope for success (link between anxiety and pathways)

USCOTS Research Spotlight Session

Nathan R. Todd
University of Illinois Urbana/Champaign

Title for Session

Assessing Student Perceptions of Lecture Usefulness: Application of a Preliminary Scale

Abstract

We developed a short survey instrument to assess student perceptions of the usefulness of undergraduate statistics lectures. We then assessed classroom lectures to investigate how students perceived a variety of teaching methods. We discuss the relationship of the scale with test performance, qualitative feedback, and attitudes toward statistics.

Purpose of this Handout:

To Discuss the Purpose and Method of This Classroom Research Project in Narrative Form

*Data are still in the process of being collected and will be presented at the Conference

Genesis of this Project:

It has been posited that active learning strategies such as group work help to facilitate classroom learning and are often suggested as effective teaching strategies. My original goal of this project was to examine this assertion in an empirical manner, by presenting material in lecture using a variety of active learning teaching methods. This was my first time to teach the class and I was excited to experiment with differing methods to see what resonated with my teaching style as well as what was effective for student learning.

I quickly realized that it is challenging to creatively assess what students actually learn in lecture as separate from what they learn from other parts of the class such as homework, lab time (a small section of 15 students meeting weekly with another TA), and other extra-class activities such as individual or group studying. The preliminary question for me then became: how do we meaningfully capture students' perceptions of lecture usefulness? Do students perceive lectures differently depending on presentation method or are all lectures experienced as similar? How do these perceptions of lecture relate to performance outcome? To general attitudes toward statistics? These foundational questions guided the research process.

General Method:

I was the instructor for the second half of the semester for two sections of an undergraduate introduction to psychological statistics class with an enrollment of 45 students per section. I taught identical material to these two class sections twice a week for 1.25 hours a class. This arrangement enabled me to teach the same material using different methods to the two different sections.

Student feedback indicated that students found examples very helpful in understanding the material, so I focused on ways of utilizing examples during lecture that might aid in student learning. I explored how group work, the chalk-board, PowerPoint, and in class data generation could serve as vehicles for effectively working through examples. For instance, I would present material to class A and would then work an example on the board whereas for class B the students would work through an example in small groups after the same material had been presented. The next class period I would swap conditions allowing for an examination within person across lecture as well as between classes to compare the specific teaching method.

Though the findings regarding the effectiveness of these various methods is interesting and instructive and will be presented, my focus became on how do we measure what is effective in the first place. As such, I generated a short nine item "Check on Lecture Learning" scale that attempted to assess student's perceptions of the lecture in helping them understand the three areas of statistical ***concepts***, ***computations***, and the ***general*** effectiveness of the lecture.

The scale was given at the end of class after different teaching methods had been employed and served as the "outcome" measure for the effectiveness of lecture. This measure may thus serve as a potential index of how individual student perceptions varied across lectures as well as may allow for the direct comparison of lectures and lecture topics.

As alluded to earlier, this measure picked up differences in class perceptions based on the teaching methods used and may serve as a useful tool to assess student perceptions of lecture usefulness. Preliminary evidence also suggests that this scale is related to class performance on exams. These findings along with the possible relationship between this lecture usefulness measure and *The Survey of Attitudes Toward Statistics* (SATS©, Schau, 2003) will be presented along with qualitative feedback provided by the students. The scale along with the subscale coding is presented below.

Quick Check on Learning Experience

These optional questions can give us an idea about what was and was not effective today in presenting the material. This is totally optional and will have NO impact on your grade and should take less than 2 minutes. Thank you for your time and feedback!

	Strongly Disagree			Neither disagree nor agree			Strongly agree
1. I feel that I learned something useful today.	1	2	3	4	5	6	7
2. Today's presentation helped me understand the new concepts introduced.	1	2	3	4	5	6	7
3. Today's presentation helped me to understand new statistical procedures.	1	2	3	4	5	6	7
4. Today's presentation helped me to think critically.	1	2	3	4	5	6	7
5. I feel confident that I could apply the computations presented today to future problems on the test and homework.	1	2	3	4	5	6	7
6. I feel comfortable with the computations covered today.	1	2	3	4	5	6	7
7. I want to see this style of presentation again in the future.	1	2	3	4	5	6	7
8. The new concepts presented today did not make any sense.	1	2	3	4	5	6	7
9. I could explain the basic concepts presented today to a friend.	1	2	3	4	5	6	7

General: Items 1, 4, 7 Procedures/Computations: 3, 5, 6 Conceptual: 2, 8r, 9 (with 8 being reversed)

Developed by Nathan R. Todd