AC 2012-4399: STUDENTS' CONFIDENCE LEVELS IN TECHNICAL CONCEPT KNOWLEDGE WITH MODEL ELICITING ACTIVITIES

Ms. Nora Siewiorek, University of Pittsburgh

Nora Siewiorek is a graduate student in the Administrative and Policy Studies Department in the School of Education at the University of Pittsburgh, where she also received her M.S. in information science. Her research interests include engineering education and educational assessment and evaluation. Her K-12 outreach activities are organizing a local science fair and a hands on workshop in nanotechnology. Her other research interests are higher education administration, comparative, and international education.

Dr. Larry J. Shuman, University of Pittsburgh

Larry J. Shuman is Senior Associate Dean for Academic Affairs and professor of industrial engineering at the Swanson School of Engineering, University of Pittsburgh. His research focuses on improving the engineering education experience with an emphasis on assessment of design and problem-solving, and the study of the ethical behavior of engineers and engineering managers. A former Senior Editor of the Journal of Engineering Education, Shuman is the Founding Editor of Advances in Engineering Education. He has published widely in engineering education literature, and is co-author of Engineering Ethics: Balancing Cost, Schedule and Risk - Lessons Learned from the Space Shuttle (Cambridge University Press). He received his Ph.D. from the Johns Hopkins University in operations research and a B.S.E.E. from the University of Cincinnati. Shuman is an ASEE Fellow.

Dr. Mary E. Besterfield-Sacre, University of Pittsburgh

Mary Besterfield-Sacre is an Associate Professor and Fulton C. Noss Faculty Fellow in industrial engineering. She is the Director for the new Engineering Education Research Center (EERC) in the Swanson School of Engineering, and serves as a Center Associate for the Learning Research and Development Center at the University of Pittsburgh. Her principal research is in engineering assessment, which has been funded by the NSF, Department of Education, Sloan Foundation, Engineering Information Foundation, and the NCIIA. Besterfield-Sacre's current research focuses on three distinct but highly correlated areas pf innovative design, entrepreneurship, and modeling. She is an Associate Editor for the AEE Journal.

Dr. Natasa S. Vidic, University of Pittsburgh
Dr. Karen M. Bursic, University of Pittsburgh

Karen M. Bursic is an Assistant Professor and the Undergraduate Program Director for industrial engineering at the University of Pittsburgh. She received her B.S., M.S., and Ph.D. degrees in industrial engineering from the University of Pittsburgh. Prior to joining the department, she worked as a Senior Consultant for Ernst and Young and as an Industrial Engineer for General Motors Corporation. She teaches undergraduate courses in engineering economics, engineering management, and probability and statistics in industrial engineering as well as engineering computing in the freshman engineering program. Bursic has done research and published work in the areas of engineering and project management and engineering education. She is a member of IIE and ASEE and is a registered Professional Engineer in the state of Pennsylvania.

Scott Streiner, University of Pittsburgh
Mr. Jeffrey Coull

Students' Confidence Levels in Technical Concept Knowledge with Model Eliciting Activities

Abstract

Assessing engineering students' technical knowledge is an important concern in engineering education. We suggest that one way to do this is by using concept inventories. These have been developed and standardized for an increasing number of the disciplines. As part of a larger NSF funded project focused on developing, incorporating and assessing Model Eliciting Activities (MEAs), we have turned to concept inventories (CI) in general, and a subset of the CI - knowledge tests (KT) - as a means of assessing conceptual understanding. The knowledge tests are focused on the particular concepts that are designed into the MEA. Included in the knowledge tests is a self-assessment of the student's level of confidence in answering each question. We are interested in studying the relative effect that MEAs designed around specific concepts can have on student learning compared to more traditional instructional methods. Although MEAs were originally designed to improve the understanding of technical concepts, our earlier research has found that they do improve students' problem solving and professional skills and result in significant learning gains; however, they may only marginally increase students' level of knowledge compared to more traditional methods. In this paper we provide an in-depth investigation of how measuring both students' performance as well as their confidence is affected by exposure to MEAs. Specifically, we ask the following: is there a significant gain in students' knowledge from the beginning to the end of the semester; are students who are most confident in their answers also correct in their responses; is there a gender difference; and, do differences exist between sections that used MEAs versus those that did not?

Introduction

Assessing engineering students' technical knowledge remains one of the most challenging concerns in engineering education. In order to properly assess students' level of understanding researchers have developed a variety of different testing instruments. One important result is the development and standardization of concept inventories for an increasing number of engineering disciplines that has occurred over the past dozen years. As part of a larger NSF funded project focused on developing, incorporating and assessing Model Eliciting Activities (MEAs), we examine how students performed on a more focused concept inventory (CI), which we have termed a knowledge test (KT) - a subset of questions focused on specific concepts that have been extracted from standardized concept inventories. In particular, we have created a KT by taking

items from two established statistics concept inventories[5, 14, 15] we have also included a self-assessment of the student's level of confidence in answering each question.

In particular, we have introduced MEAs into an Introductory Engineering Statistics course to see if they might improve conceptual understanding more effectively compared to traditional instruction methods that did not utilize the MEAs. To measure this, as noted we created a knowledge test from two statistics concept inventories. Although MEAs were originally designed to improve the understanding of technical concepts, our research has found that MEAs may only marginally increase students' level of knowledge compared to these more traditional methods; however they do result in substantial improvements to students' problem solving and professional skills compared to the methods more typically used in engineering classrooms[12, 13].

In this paper we provide an in-depth investigation of how both students' performance as well as their confidence relative to a few key concepts is affected by exposure to MEAs.

Research Questions

This study aims to address the following five research questions.

1. Is there a significant gain in students' knowledge from the beginning to the end of the semester?
2. Are students who are most confident in their answers also correct in their responses?
3. Are there differences in confidence related to gender?
4. Do differences exist between experimental and comparison sections?
5. Are there differences simply due to misunderstandings, or are these more likely long held misconceptions?

Background

1. Challenges for Students Learning Statistical Concepts

Historically students rarely encountered statistical concepts before the college level; however efforts by the National Council of Teachers of Mathematics in the 1980's encouraged teaching statistics at the K-12 grades[1]. These educators initially focused on how to teach statistics, but soon realized that they had to examine on how students learn if they were going to be effective. Two of these researchers – Garfield and Ahlgren - have noted that "The experience of psychologists, educators, and statisticians alike is that a large proportion of students, even in college, do not understand many of the basic statistical concepts they have studied." As a result they found that students often default into "number crunching mode," and have other ways of thinking or incorrect intuitions, which contribute to the challenge of students learning statistical concepts. They point out that "Recently, some research on problem solving has shown that students receiving deliberate instruction in how to solve problems do become better problem solvers and are better able to "think mathematically"[1]. Garfield[2] has outlined seven theories of learning and certain principles to help students learn statistics, noting the importance of feedback after testing, and identifying and understanding students' misconceptions, which we suggest that MEAs combined with concept inventories or knowledge tests might be able to accomplish. Garfield points out that while students can do math calculations and understand the basic

concepts, they are challenged when they have to transfer these concepts to real world applications[2].

Development and use of Concept Inventories in Engineering

Concept Inventories have been developed for a variety of engineering and science areas. The literature now includes studies not only on the writing and development of CIs, but also how to use and implement them as well[3, 4, 5]. It is now generally accepted that concept inventories can be used to help identify which technical concepts are the more challenging for students to grasp, and provide some indication of the overall success in gaining conceptual understanding from a particular course or series of courses. Given at the start of the course (pre-test), they provide the instructor with an overview of the students' initial knowledge level; such information enables the instructor to more effectively tailor the course to put appropriate emphasis on those concepts that students do not yet grasp, while devoting less effort on those concepts that have already been learned. The pre-test concept inventory also allows the instructor to identify if there are students who are either way above or way below the general conceptual understanding of the class. This should be followed by a post-test concept inventory, administered at the end of the term, which then allows the instructor to better calibrate the extent of student learning, and, as appropriate revise the course material and syllabus for the next time it is taught. If the course is the first in a sequence, or is a foundation course, providing the instructor(s) of the follow-on course(s) with this information should further enhance learning.

Yet, with all the effort that has gone into concept inventories, to date, there is not a commonly accepted definition of what they are, nor is there one agreed on method for developing them[7]. However, some common aspects of concept inventories have emerged. By necessity, they are almost all multiple-choice instruments. In addition to the correct answer, a well-designed CI will use distractors as possible answer choices. These distractors are typically determined through pilot testing and cognitive interviews (i.e., verbal protocols) where the researchers identify typical responses for students who have not fully grasped the concept, either due to misunderstanding, or, more seriously, misconceptions. For a well-designed concept inventory, selection of particular incorrect answers provides the instructor with an indication if the student completely guessed the answer, had only a partial or incomplete understanding (i.e., misunderstanding), or, much more seriously – a misconception. While misunderstandings are correctable, misconceptions are much more challenging to repair.

Misconception vs. Misunderstanding

In efforts to understand students reasoning about statistical concepts, Kahneman and Tversky's "representativeness"[1,17], researchers have investigated statistical misconceptions[17]. The term misconception is often used more broadly to cover preconception, misunderstanding, misuse, or misinterpretation. Misconception can also be defined more narrowly. Misconception is a long held belief that a student has prior to starting the course. For example, what does "variation" mean? Additional statistical topics where misconceptions occur include sampling variability, significance levels, and statistical significance[18].

In contrast to a misconception, a situation when a student does not correctly understand the concept in the course is defined as misunderstanding. Brewer's study of statistical textbooks

outlines five "myths and misconceptions" about statistics: half-truths, definitional errors, constant-cum-variable, cart-before-the-horse and unitary inference. Brewer also investigates which statistical concepts misconceptions are often about: hypothesis testing, confidence intervals, sampling distributions and the central limit theorem[19].

Reed-Rhoads and Imbrie note that:

> "Great care goes into conceptualizing the nature of the situations to be presented and in developing plausible distractors that represent a range of partially correct understandings to completely incorrect understandings and misconceptions. . . . In recent years, the science, technology, engineering, and mathematics (STEM) disciplines have increased their use of Concept Inventories (CI) instruments to measure the value added to student learning by new ways of teaching important material. Utilizing a tool such as a CI can provide a learning opportunity for students and professors alike."[7].

Two websites that provide information about CIs, including a number of these instruments are:

- The Concept Inventory Hub, or CI Hub at http://dev.cihub.org/, a community for concept inventory developers, researchers, faculty and students.
- Concept Inventory Central (CIC), https://engineering.purdue.edu/SCI/workshop/tools.html, which has links to concept inventories in many engineering topics, in addition to math and science disciplines.

2. Development and use of Model Eliciting Activities (MEAs) in Engineering

MEAs present complex, realistic, open-ended problems to students to reinforce targeted concepts[8-12]. Students solve these realistic, client driven problems in teams and are incorporated into an existing course structure as either a project or homework assignment. MEAs are designed according to six principles as scaffolding for students to either: integrate, reinforce or discover new concepts[8, 10]. We have extended the MEA methodology by introducing an ethical dimension that students must consider in the problem scenario[9]. The student teams must report their proposed general "model" and specific solution in memo format to a fictitious client. By requiring the team to develop and report out a generalizable procedure, the MEA construct helps to reveal their thought processes (including assumptions, decisions made about the problem and solution strategies).

Because MEAs can address a combination of technical and professional skills, it is more challenging to assess the resultant student learning. For full impact, instructors must guide the students' learning and provide targeted feedback; especially if it appears that misconceptions exist, often best observed through students' self-reflection reporting. We have reported elsewhere the impact of MEAs on conceptual learning and the instructors' perspectives about using MEAs in the classroom[11], and the improvement of student attainment of ABET outcomes, especially the professional skills, in using MEAs[12, 13].

Knowledge Tests and MEAs in an Introductory Engineering Statistics Course

For this particular study, rather than use the full statistics concept inventory developed by Allen et al.[14], we carefully selected questions in order to create a knowledge test (i.e., focused CI). We were interested primarily in learning more about student attainment of a subset of basic concepts embedded in our MEAs, rather than the full spectrum of concepts in the complete inventory.

The focused concept inventory (KT) used in this study is the set of 20 multiple choice questions selected from two established and tested concept inventories. The questions were chosen to correspond to the targeted technical concepts reinforced by the MEAs, which were specifically adapted for the course. Eighteen of the twenty questions were used (with permission) from the Statistics Concept Inventory (SCI) developed by Allen, Stone, Reed-Rhoads and Murphy[14]. The remaining two questions were used (also with permission) from the Comprehensive Assessment of Outcomes in a first Statistics Course (CAOS) developed by Garfield, delMas, Chance and Ooms[15]. The authors of the CAOS[15] have noted that the SCI was written for a more targeted engineering student audience, while the CAOS instrument is broader in both the statistical content and the range of students who take statistics courses.

As noted, in addition to examining how each student performed on this focused concept inventory test, we also included his/her level of confidence in each answer. This was consistent with prior work[16] by Reed-Rhodes and her colleagues (SCI authors) in order to discern if students were simply guessing or did have incorrect understanding about a concept. Students rated their confidence for each response on a four point scale, ranging from 0 (complete guess) to 3 (very confident). See appendix for the selected set of questions from the CI used in this study.

Table 1: MEAs used in the Introductory Statistics course

MEA Title	Decision Situation	Ethical Dilemma	Targeted Technical Concepts
Tire Reliability	Develop a general procedure to analyze reliability of any set of tires based on "acceptable reliability" data set	Safety concerns about reliability of a tire production run	reliability, mean, median, standard deviation, histogram, probability plots, percentage, outliers
Test Leads	Develop a sampling procedure to ensure a batch of test leads is acceptable dimensions, including the minimum sample size for the expensive product	Determining conditions under which a recall of defibrillators might be recalled	central limit theorem, uniform distribution, sample size, means, sample of the means, confidence intervals, variance, sampling distribution
CNC Machine	Comparing the performance of two types of machines to determine if a new machine should be purchased	Determining the weight of management's advice and reporting realistic results	hypothesis testing, standard deviation, confidence intervals, variance, central tendency

The MEAs were carefully introduced into the introductory engineering statistics course through a quasi-experimental design in which there were both experimental and comparison sections[12]. To date there have been four experimental courses in which the MEAs have been introduced and five comparison sections in which they were not used. The course – Probability and Statistics for Engineers 1 is a required core course for industrial engineering students, typically taken in the sophomore year. Additional sections of the course are offered to the other engineering

disciplines, primarily as a requirement, although a few students take it as a technical elective; these students may take the course as sophomores, juniors or seniors.

A brief overview of each MEA and the targeted technical concepts embedded in each one is included in Table 1. For additional examples of MEAs as well as guidelines for their use please see http://modelsandmodeling.net/Home.html.

Methodology

For this study, we compared two experimental sections (ES-1 and ES-2) and three comparison sections (CS-1, CS-2 and CS-3). The focused concept inventory was administered at the beginning of the term and at the end of the term in all five sections. The same instructor taught both experimental sections; the three comparison sections were each taught by a different instructor. The instructor using MEAs (ES sections) was experienced in using this construct within the classroom setting. The same content was covered in all sections including traditional homework assignments, quizzes and three exams during the semester. The only difference was the implementation of MEAs in experimental sections[12]. The Fall 2010 experimental section consisted of industrial engineering (IE) students; the Spring 2011 experimental section and the three comparison sections all consisted of students from across the engineering school. The detailed methodology is presented in Vidic, et al[12]. The number of students enrolled and the term offered for each section is presented in Table 2. (It should be noted that when the KT was first used in Fall 2010 two of the questions were repeated. This was corrected for the Spring 2011 CI. Therefore, the Fall 2010 CI is out of 18 questions and the Spring 2011 version has 20 questions.)

Table 2: Different Course Sections and enrollment

Section of the Course	Term offered	Number of Students
ES-1: MEA section, instructor 1	Fall 2010	58
ES-2: MEA section, instructor 1	Spring 2011	52
CS-1: Non-MEA section, instructor 2	Fall 2010	71
CS-2: Non-MEA section, instructor 3	Fall 2010	69
CS-3: Non-MEA section, instructor 4	Spring 2011	62

As noted, a level of confidence question was asked after each concept question to finesse whether or not the student felt that he or she understood the concept, was simply guessing, or had a misunderstanding or misconception about the technical concept (i.e., indicated high confidence level for an incorrect answer). Specifically, we asked for each item:

0- I feel clueless about the answer
1- I think this might be the right answer
2- I feel pretty good about the answer
3- I am completely sure it is right

For each student, his or her score on the pre and post CI was recorded, in addition to final course grade and demographic information such as course year, engineering major and gender.

Results

1. Is there a significant gain in students' knowledge from the beginning to the end of the semester?

Analysis of student responses to the focused concept inventory test shows increased learning gains from pre to post for all five sections. There is a significant statistical difference (p-value $\leq$ 0.0001) between the average KT scores at the beginning and end of the term for all three groups.

However, there was not a significant difference between experimental and comparison sections during the Fall 2010 semester. There was the statistical difference in the mean score between the experimental and comparison sections in the Spring of 2011 (*p-value=0.012; one tailed test).* For Fall 2010, the effect sizes were large for experimental section and two of the comparison sections. For Spring 2011, comparison section had a medium effect size; however, that section had the highest pre-test score of all five sections. That is, students started at a higher level but gained less than those who started at lower levels. The KT scores as well as the effect sizes are presented in Table 3. Note that students tended to answer an average of three more questions correctly on the post-test, compared to the pre-test, which is where the gain in conceptual understanding was achieved. The two experimental sections did have higher post scores by 0.44 and 1.56 compared to the highest scoring comparison section.

Table 3: KT scores for the Fall 2010 and Spring 2011

	Item	ES-1	CS-1	CS-2	CS-3	ES-2
Start Term	**Mean**	**7.21**	**6.91**	**6.81**	**8.26**	**8.04**
	St. Dev.	2.62	2.67	2.82	2.08	3.03
	Sample Size	58	70	68	47	46
End Term	**Mean**	**10.56**	**10.12**	**10.02**	**10.07**	**11.68**
	St. Dev.	2.61	2.83	2.72	3.09	3.29
	Sample Size	57	66	65	41	41
Effect size		**1.28 Large**	**1.15 Large**	**1.17 Large**	0.70 Medium	**1.15 Large**

2. Are students who are most confident in their answers also correct in their responses?

In addition to gain in their conceptual knowledge we wanted to determine if there was any difference in the levels of confidence for students in experimental sections compared to those in comparison sections.

We calculated the average confidence levels for the pre-test and the post-test as well as the average confidence per correct answer and the average confidence per incorrect answer for the post test for each section. The average confidence levels increased from pre to post in all course sections. These scores are presented in Table 4.

As expected, for each section student confidence for the correct responses were significantly higher than confidence on incorrect responses. (*Paired-t; p<.001*). Confidence for correct answers compared to incorrect ranged from a third to a half of a point higher. This tends to suggest, that students were not guessing, especially for those questions that they were able to answer correctly on the post-test.

Table 4: Average confidence scores

	ES-1	**ES-2**	**CS-1**	**CS-2**	**CS-3**
Pre average confidence	1.14	0.95	1.12	1.15	1.09
Post average confidence	2.04	1.91	1.99	1.88	1.85
Confidence per correct answer post	2.23	2.03	2.16	2.05	2.01
Confidence per incorrect answer post	1.76	1.70	1.78	1.67	1.64
Probability Paired t	<.001	<.001	<.001	<.001	<.001

We also tried to determine if there is a linear relationship between the total score and the level of confidence for each student. Using our sample of 270 students, a very small relationship was found ($R^2 = 0.09$). We also looked for a relationship between average confidence for those items answered correctly compared to the percent answered correctly. We found a positive, but very weak trend ($R^2 = 0.07$); that is, as the number of correct responses increases, the average confidence also increases.

To better understand what was occurring, we analyzed the relationship between the percentage of students who answered the item correctly on the post-test and their average confidence. As we expected, in general the more students who answered the question correctly, the higher their average confidence score; $R^2 = 0.69$ (i.e., 69% of the variation explained), as shown in Figure 1. (Note that the highest confidence score is 3.) In contrast, if we examine the correlation between the percentage of students who answered the item wrong and their average confidence, we see a

negative relationship, again as would be expected. In this case, $R^2 = 0.24$ (24% of the variation explained). This suggests that as students better understand a concept their confidence increases. Conversely, the less the understanding, the less is their confidence. Note that Figure 1 is a plot of each of the 20 items/questions on the concept inventory (adjusted for having only 18 items for the Fall 2010 section).

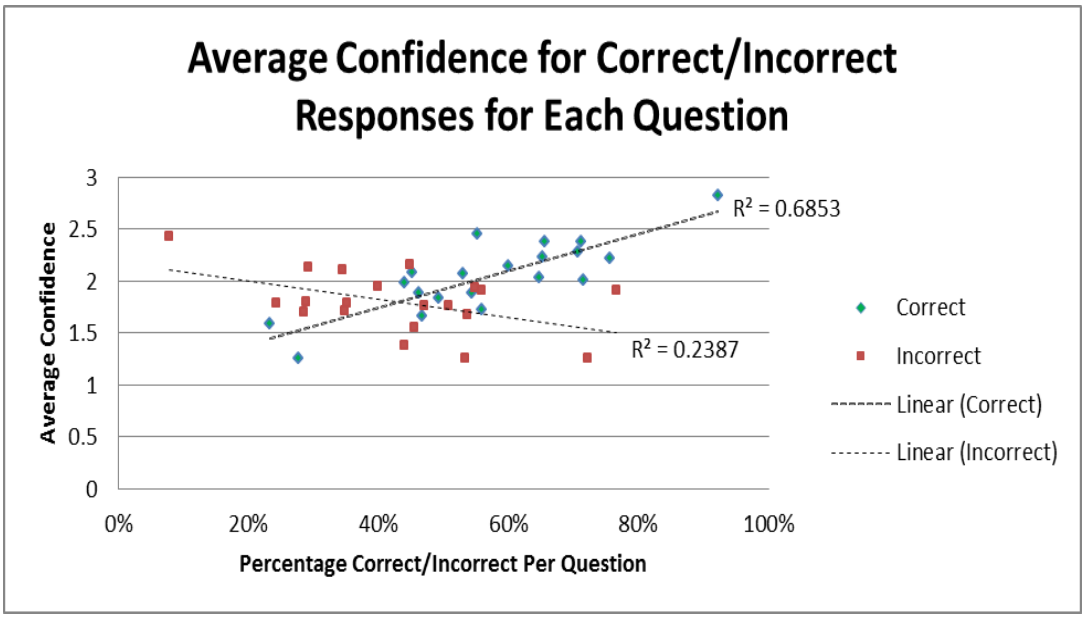


Figure 1: Average Confidence for both correct and incorrect answers for the concept inventory

The average post-test score by grade earned in the course for each of the five sections is presented in Table 5. Not surprisingly, it shows relative consistency in terms of improving concept score with higher grade. It also gives the average confidence, which suggests a U-shape curve. That is, those earning an A tended to have the highest confidence, with slight decreases for B and then C grades. However, those students who earned a D or failed the course in some cases displayed the same confidence in their concept inventory responses as those getting the best grades, even though many more of their responses were wrong.

Table 5: Average Post KT Score and Confidence per section and Course Grade

Section	A		B		C		D		F	
	Score	Conf.	Score	Conf.	Score	Conf.	Score	Conf.	Score	Conf.
ES-1	12.00	2.17	10.00	1.94	8.67	1.56	7.75	2.07	7.67	2.02
ES-2	13.38	1.96	11.00	1.94	9.60	1.69	8.00	0.95	NA	NA
CS-1	11.10	2.04	10.30	2.02	8.08	1.85	7.00	2.22	6.00	1.67
CS-2	11.69	1.96	9.27	1.85	9.82	1.81	7.00	1.68	5.67	1.44
CS-3	12.80	2.03	9.87	1.82	9.33	1.36	5.50	1.55	7.00	2.18

Table 6 illustrates this situation more clearly. In this case we have divided the average confidence of all of the student's responses by the number of correct answers. Given the relatively low level of variability in confidence, one would then expect that those with few correct answers would exhibit the higher ratios, which is what the table indicates for those

earning a D or failing the course. In contrast, the ratio is relatively constant for those earning an A, B or C, consistent with comparable decreases in both score and confidence. There is also no apparent difference between the experimental and comparison sections; both exhibit similar patterns.

Table 6: Average confidence per correct answer

Section	A	B	C	D	F
ES-1	0.18	0.19	0.18	0.27	0.26
ES-2	0.15	0.18	0.18	0.12	NA
CS-1	0.18	0.20	0.23	0.32	0.28
CS-2	0.17	0.20	0.18	0.24	0.25
CS-3	0.16	0.18	0.15	0.28	0.31

3. Are there differences in confidence related to gender?

Table 7 summaries the average confidence by gender. Note that in all cases, the male students' confidence per correct answer, as well as incorrect answer is higher than that for the female students, as what might be expected. However, the only significant difference (*p value=0.025*) is for the second experimental section for the male average correct confidence of 2.11 compared to the female average of 1.64. Also note that in all cases, those selecting the correct answer were more confident than those selecting an incorrect answer.

Table 7: Summary by Gender

Female					
Section	**Pre**	**Post**	**Conf.**	**Conf. Right**	**Conf. Wrong**
ES-1	6.20	10.37	1.93	2.08	1.73
ES-2	7.92	10.33	1.67	1.64	1.58
CS-1	6.64	8.93	1.90	2.08	1.71
CS-2	6.11	10.23	1.77	1.95	1.54
CS-3	8.09	9.82	1.57	1.78	1.57
Male					
Section	**Pre**	**Post**	**Conf.**	**Conf. Right**	**Conf. Wrong**
ES-1	7.35	10.66	2.09	2.21	1.89
ES-2	8.09	12.24	1.98	2.11	1.76
CS-1	6.98	10.34	2.02	2.15	1.88
CS-2	7.09	10.08	1.93	2.05	1.73
CS-3	8.29	10.17	1.85	1.97	1.69

We found similar relationships between the student's overall confidence and his or her overall percent correct. For both male and female, and positive slope was found; however, only nine percent of the variation was explained - a significant, but relatively weak relationship. Further, for both male and female responders a similar positive, but weak relationship between average

confidence and percent of questions answered correctly was found; in both cases, only six percent of the variation is explained. When average confidence for incorrect answers is examined, there is no relationship observed for the female students, and a very small negative relationship for males as the percent of questions answered wrong increases. Note that one might expect to see a negative relationship; i.e., as the number of incorrect responses increases, the student's average confidence would decrease.

4. Do differences exist between experimental and comparison sections?

Shown in Table 8 below are the percent correct for each individual KT question per section of the course. We have arbitrarily divided the scores on the concept inventory into three divisions: very good or above 70%, average or between 50%-70% and poor below 50%. Note that a higher proportion of the students in the experimental sections scored in the top (very good) category, compared to the comparison section.

Table 8: Summary of Correct Responses

Sector	ES-1	ES-2	CS-1	CS-2	CS-3
More than 70% correct	38.9%	30%	22.2%	22.2%	5%
50% to 70% correct	33.3%	35%	38.9%	44.4%	45%
Less than 50% correct	27.8%	35%	38.9%	33.3%	45%

5. Can differences be attributed to long held misconceptions or simply misunderstandings?

In order to better determine whether or not students had long held misconceptions, or it was more a misunderstanding, we investigated four of the 20 CI questions in closer detail (these questions can be found in the appendix). We wanted to better understand how and why students might be responding the way they did and what it indicated about their conceptual understanding. The four selected were the ones where the overall student performance was the weakest. Interestingly, two of these items involved the student interpreting plots (questions 7 and 11), and two involved the t-distribution and/or t-test (questions 15 and 18). The percent answering these questions correctly was 24%, 45%, 15% and 36% respectively.

Table 9 summarizes the results for these four questions. A 2 or 3 (I feel pretty good about the answer or am completely sure it is right) is designated as a high response; a low response is 0 or 1 (I feel clueless or I think this might be the right answer).

For one of these four questions, the average confidence was higher for the incorrect answers compared to the correct answers. For the other three, it was about the same, although for one the confidence for both correct and incorrect answers was very low at 1.26 and 1.25 respectively, suggesting that students had some idea of the correct answer, enabling them to at least eliminate one of the choices, but guessing at the other three (i.e., just over a third selected the correct answer). Further, for three of the four questions, over two-thirds of the students were considered to be highly confident about these answers; for only the question on the t-test did almost two-

third of the students exhibit low confidence. Further, for those who were highly confident of their response, in two cases, at best they were only correct about half the time, but in the other two they were at best correct 1/5 and 1/7 times respectively.

Question 7 required students to apply the central limit theorem in interpreting a set of plots. The central limit theorem was the primary concept that was reinforced by the second MEA. Students were given a plot of the density function and told that 10 random data points were drawn from that function and the mean computed. This was repeated 20 times. The observed means were then placed into a histogram with six bins. Students were give four different plots and told to select the correct one.

As noted, only 23.33% selected the correct answer, while three times that number of students selected the primary distractor – a distribution that was similar to the shape of the density function. The experimental section actually did slightly poorer than the comparison section and exhibited lower confidence. Further, those choosing the distractor were the most confident of their selection. These results suggest misunderstanding of the basic concepts of both the central limit theorem and sampling.

Table 9: Four most "difficult questions"

Item	Confidence	Correct Response	Incorrect Response
Question 7	Low (31.8%)	27 (31.8%)	58 (67.2%)
Sampling from give distribution; construct a histogram; select correct plot	High (67.2%)	36 (19.8%)	146 (80.2%)
Average Confidence		*1.59*	*1.91*
Question 11	Low (24.0%)	20 (31.3%)	44 (68.7%)
Given four histograms; which one shows the most variability	High (76.0%)	100 (49.3%)	103 (50.7%)
Average Confidence		*2.08*	*1.93*
Question 15	Low (24.4%)	12 (18.5%)	53 (81.5%)
Given four statements; select which ones are correct for the t-distribution	High (75.6%)	29 (14.4%)	172 (85.6%)
Average Confidence		*1.99*	*1.92*
Question 18	Low (62.9%)	50 (29.8%)	118 (70.2%)
Give a t-test and four statements, which ones are correct for the particular test	High (37.1%)	47 (47.5%)	52 (52.5%)
Average Confidence		*1.25*	*1.26*

Question 11 involved four histograms; the students were asked to select the one with the most variability. One of the histograms resembled a uniform distribution (correct answer); a second had a sharp mean, and rather uneven spread (first distractor) a third resembled a normal

distribution (second distractor). The fourth was also uniform, but with a smaller range. In developing the index Allen and his colleagues found that the first distractor was always the most common answer[20]. Their focus groups indicated that it is often chosen because it is the "bumpiest." They also found that the second distractor was also frequently selected because of the "normal shape" and "they are familiar with it". They noted that many students did not interpret the histogram correctly, but read it instead as actual scores; i.e., as if it were a bar chart (non-frequency) or scatter plot rather than frequency counts.

Table 10: Results for Question 7

Alternative	Percent Selected	Average Conf.	ES Selected	ES Conf.	CS Selected	CS Conf.
Correct	23.33%	1.59	20.6%	1.30	25.2%	1.58
Distractor	68.89%	1.95	68.0%	1.86	69.6%	1.97
Alterative 1	4.81%	1.62	7.0%	1.29	2.9%	2.00
Alternative 2	2.22%	1.33	4.0%	2.00	2.3%	1.00

As shown in Table 9, the experimental students selected the first distractor more frequently than the correct response, while the comparison students actually selected the correct response more frequently. Both groups made these selections with relatively high confidence. That was not true for the second distractor, which was selected by 1/11 of the experimental group and 1/8 of the comparison. The small number of students who selected the alternative did so with little confidence and was most likely guessing at the answer.

Question 15 involves four statements about the t-distribution:

a. It is used for small samples
b. It is used when the population standard deviation is not known
c. It has the same basic shape as a normal distribution but has less area in the tails

Students are asked to select which ones are correct – a, b, c, a and b (correct), or all three (distractor).

Table 11: Results for Question 11

Alternative	Percent Selected	Average Conf.	ES Selected	ES Conf.	CS. Selected	Comp. Conf.
Correct	45.19%	2.08	41.84%	2.05	47.09%	2.10
Distractor 1	42.22%	1.97	48.98%	2.04	38.37%	1.91
Distractor 2	11.11%	1.90	9.18%	1.78	12.21%	1.95
Alternative 1	1.48%	1.25	0		2.33%	1.25

Both the experimental and comparison sections students did approximately the same on this question. Students who selected either the correct answer or the first distractor were equally, highly confident. However, less than half of the students got the correct answer for this question.

Table 12: Results for Question 15

Alternative	Percent Selected	Average Conf.	ES Selected	ES Conf.	CS Selected	Comp Conf.
Correct	44.07%	1.99	45.9%	1.87	43.0%	2.00
Distractor 1	37.04%	1.97	37.8%	1.92	36.6%	1.99
Alternative 1 (a)	15.93%	1.88	14.3%	1.79	16.9%	1.50
Alternative 2 (b)	1.11%	1.67	1.0%	2.00	1.2%	1.25
Alternative 3 (c)	1.85%	1.40	1.0%	2.00	2.3%	2.07

Question 18 involves a one-tail test of hypotheses, one of the latter topics covered in the course, but one that the third MEA is built around. The test is rejected with a p-value of 0.10. The students are given four statements and asked which one is correct:

a. The test statistic fell within the rejection region at the $\alpha = 0.05$ significance level (distractor)
b. The power of the test statistic used was 90%
c. Assuming H_0 is true, there is a 10% possibility that the observed value is due to chance (correct)
d. The probability that the null hypothesis is not true is 0.10

Stone[20] reports that the response distribution indicates widespread confusion about p-value and hypothesis testing. They note that the first choice has been the most attractive, and that the number of correct responses almost always falls after instruction, indicating misunderstanding. The results for both the experimental and comparison groups are similar to what Allen, et al. report[20]. In fact, the comparison group did slightly better than the experimental, again, consistent with the developers' findings. Confidence is relatively low for all choices, indicating that students are not very sure about their responses and, may, in fact be guessing. Again, this suggests misunderstanding of the basic concepts that were first introduced in the course.

Table 13: Results for Question 18

Alternative	Percent Selected	Average Conf.	ES Selected	ES Conf.	CS Selected	Comp Conf.
Correct	27.78%	1.25	23.4%	1.17	30.0%	1.29
Distractor 1	37.04%	1.45	43.9%	1.56	33.3%	1.37
Alternative 1	14.81%	0.93	17.3%	0.65	13.5%	1.13
Alternative 2	19.63%	1.15	15.3%	1.53	22.2%	1.00

Discussion

Additional factors such as course grade, engineering major, class year and gender were also examined. We did find a correlation between concept inventory test score and the student's final grade in the course. We also found a correlation between students' confidence scores and final course grades, something we had not observed in the literature[3]. In contrast, major and grade level was not statistically significant in any test that was carried out. Results suggest that these factors do not affect course grade, KT score, or even average confidence with any significance.

While we found significant learning gain in both the experimental and comparison sections as measured by the knowledge test, which is based on a subset of proven concept inventory questions, one major concern remains – at best all sections were on average only getting approximately half of the twenty items correct on the post-test. As noted, the post-test came at the end of the course. The concepts contained in the KT were covered in the course, and in the case of the experimental sections, were also covered in the MEAs. Stated another way, the pre-test scores ranged from a low of 6.81 to a high of 8.26 for the five sections; however, the post-scores only ranged from 10.02 to 11.26. Thus the students on the pre-test averaged 7.5 items correct, and after taking the course, got another 3 items correct.

We had originally turned to the MEAs as a potential tool to improve conceptual understanding. While the use of MEAs in these instances did improve conceptual understanding significantly, this improvement was no different than what resulted from the more traditional teaching methods that didn't use MEAs. In both cases, as discussed above, both resulted in students on average getting an additional three questions correct, but they still got almost half of the targeted concept questions wrong. Hence, the challenge remains for engineering educators – how do we increase conceptual understanding? Does this suggest a significant amount of misunderstanding, which ideally should be correctable, or does it really suggest that students enter with misconceptions, which are extremely difficult to repair. If the latter is correct, then it is not surprising that students still miss half of the conceptual items after completing the course.

Conclusions

We posed five questions at the beginning of this paper:

First – is there a significant gain in students' conceptual knowledge from the beginning to the end of an engineering statistics course? Here the answer is a resounding "yes"! Using our knowledge test with twenty items extracted from two confidence inventories, we found very high levels of significance between pre and post tests in experimental as well as comparison sections. However, the experimental sections had highest scores on the post tests as well as largest gains.

Second – are students who have the most correct answers also exhibit the highest confidence? Here again, the answer is yes, but not as definitive. There is still a lot of "noise," so that relationships are not as strong as one might expect. So the more questions that a student answers correctly, the higher his or her confidence. Further, as the number of correct answers for a given question increases, so does the confidence of those answering correctly.

Third – are there any differences due to gender? Here, the answer is "not really." While we did find that male students tended to answer questions with higher confidence than female students, the differences for the most part were not significant. This, in itself, is an encouraging result. We would hope with upper level students to not find significant differences in confidence.

Fourth – do differences in KT scores and confidence exist between our experimental and comparison groups? Here the answer is a more nuanced one – we found a higher proportion of students in the experimental sections tended to score better than for the comparison sections. Yet, it was difficult to find statistical differences when comparisons were made using other measures. What we did learn from other studies is where our experimental groups stood out (i.e., where the MEAs seemed to have the most positive impact) was in acquiring the ABET professional skills, certainly something to be taken seriously.

Fifth, and finally – to what extent can these differences be attributed to misunderstanding rather than misconceptions? In examining four of the more difficult questions in greater detail, we the data suggests that misunderstanding may exist to a greater extent than we might have originally anticipated and, while the MEAs are a good construct for enabling students to learn to master the professional skills, we have yet to document that they can improve conceptual understanding relative to more traditional methods of instruction.

Acknowledgment

This research is supported in part by the National Science Foundation through DUE 071780: "Collaborative Research: Improving Engineering Students' Learning Strategies through Models and Modeling." We also thank Dr. Teri Reed-Rhodes (Purdue University) for suggesting that much of what we had called misconceptions were really misunderstandings.

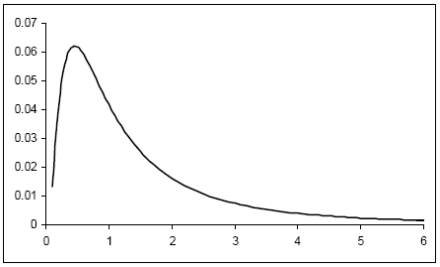
References

1. Garfield, J. and 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.
2. Garfield, J. (1995). "How Students Learn Statistics", *International Statistical Review*, 63(1), 25-34.
3. Steif, P.S. and Hansen, M.A. (2007). "New Practices for Administering and Analyzing the Results of Concept Inventories", *Journal of Engineering Education*, 96(3), 205-212.
4. Stone, A., Reed-Rhoads, T., Murphy, T.J., and Imbrie, P.K. "Use of Item Response Theory to Facilitate Concept Inventory Development", Proceedings of the Research in Engineering Education Symposium 2009, Plam Cove, QLD.
5. Stone, A., Allen, K., Reed-Rhoads, T., Murphy, T.J., Shehab, R., and Saha, C. "The Statistics Concept Inventory: A Pilot Study," Frontiers in Education Conference, Boulder, CO, 2003.
6. Allen, K., Reed-Rhoads, T., Terry, R.A., Murphy, T.J., and Stone, A.D. (2008). "Coefficient Alpha: An Engineer's Interpretation of Test Reliability", *Journal of Engineering Education*, 91(1), 87-94.
7. Reed-Rhoads, T. and Imbrie, P.K. "Concept Inventories in Engineering Education", NAE Commissioned Paper for the Board on Science Education, Center for Education. http://www7.nationalacademies.org/bose/Reed_Rhoads_CommissionedPaper.pdf
8. Yildirim, T.P., L.J. Shuman, and M. Besterfield-Sacre. "Model Eliciting Activities: Assessing Engineering Student Problem Solving and Skill Integration Processes." *International Journal of Engineering Education* 26, no 4 (2010): 831-845.

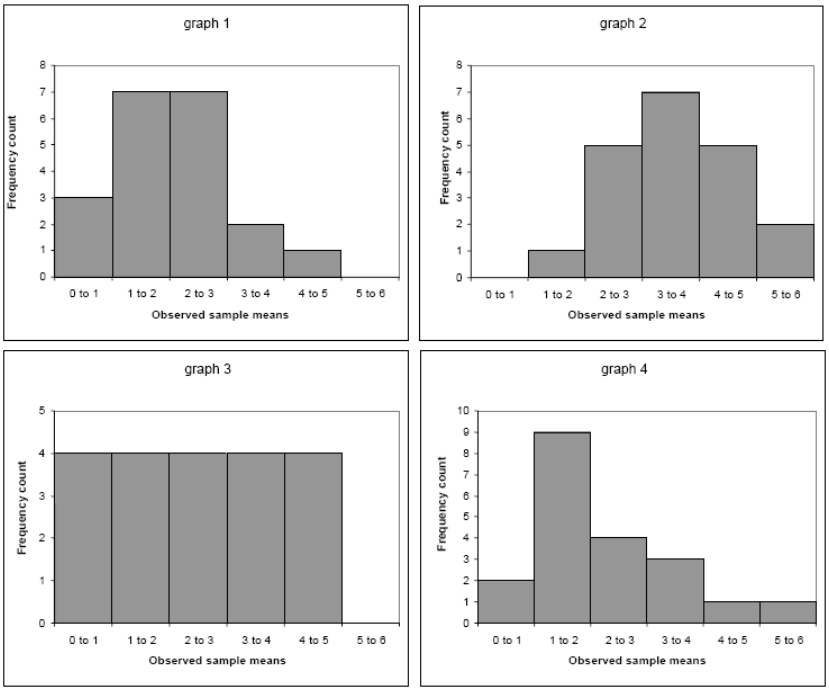
9. Shuman, L.J., Besterfield-Sacre, M., R. Clark, Yildirim, T.P. and K. Bursic (2009) "Introducing An Ethical Component to Model Eliciting Activities", *2009 American Society for Engineering Education National Conference*, Austin, TX, June 14-17, 2009.
10. Shuman, L., T. Moore, M. Besterfield-Sacre, H. Diefes-Dux, E. Hamilton, R. Miller, B. Olds, and B. Self, "Improving Engineering Students' Learning Strategies Through Models and Modeling," 38th ASEE/IEEE Frontiers in Education Conference, Saratoga Springs, NY, October 22-25, 2008.
11. Siewiorek, N., Shuman, L.J., Besterfield-Sacre, M., Goldstein, R. "Comparison of Instructor Perceptions and Student Reflections on Model Eliciting Activities", 2011 Proceedings of the American Society for Engineering Education Annual conference and Exposition, Vancouver, British Columbia.
12. Vidic, N., Shuman, L.J., Besterfield-Sacre, M., Bursic, K., Yildirim, T.P, and Siewiorek, N. "Learning Impacts Gained from Model Eliciting Activities (MEAs) in an Introductory Statistics Course", 2011 Proceedings of the Industrial Engineering Research Conference, Reno, Nevada.
13. Bursic, K., Shuman, L.J., and Besterfield-Sacre, M., "Improving Student Attainment of ABET Outcomes Using Model Eliciting Activities (MEAs)", 2011 Proceedings of the American Society for Engineering Education Annual conference and Exposition, Vancouver, British Columbia.
14. Allen, K., Stone, A., Reed-Rhoads, T. and Murphy, T.J. "The Statistics Concepts Inventory: Developing a Valid and Reliable Instrument", ASEE Conference, 2004.
15. delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). "Assessing students' conceptual understanding after a first course in statistics", *Statistics Education Research Journal*, *6*(2), 28-58.
16. Allen, K. Reed-Rhoads, T. and Terry, R. "Work in Progress: Assessing Student Confidence of Introductory Statistics Concepts", Frontiers in Education Conference, San Diego, CA, 2006.
17. Callaert, H. (2002). "Understanding Statistical Misconceptions", ICOTS6. http://www.stat.auckland.ac.nz/~iase/publications/1/10_07_ca.pdf
18. Castro Sotos, A.E., Vanhoof, S., Van den Noortgate, W., Onghena, P. (2007). "Students' Misconceptions of Statistical Inference: A Review of the Empirical Evidence from Research on Statistics Education", *Educational Research Review*, 2, 98-113.
19. Brewer, J.K. (1985). "Behavioral Statistics Textbooks: Source of Myths and Misconceptions?" *Journal of Educational Statistics*, 10(3), 252-268.
20. Stone, A. (2006). "A Psychometric Analysis of the Statistics Concept Inventory", PhD Dissertation.

Appendix

7.



From the above probability density function, 10 random data points are drawn and the mean is computed. This is repeated 20 times. The observed means were placed into six bins to construct a histogram. Which of the following histograms is most likely to be from these 20 sample means?

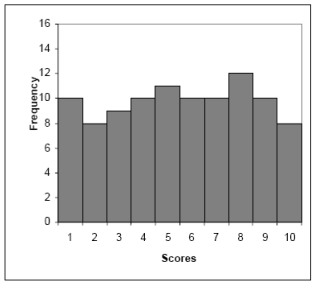


a) Graph 1
b) Graph 2
c) Graph 3
d) Graph 4

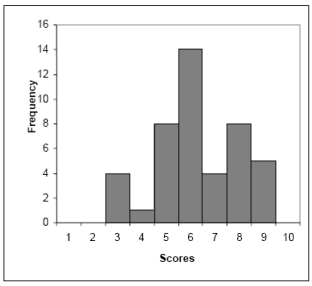
Please check the answer that describes your level of confidence in the answer you picked.

	I feel clueless about the answer.	I think this may be the right answer.	I feel pretty good about the answer.	I am completely sure it is right.
Confidence Level				

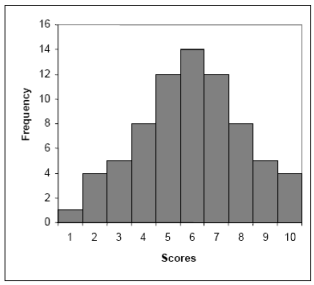
11. The following are histograms of quiz scores for four different classes. Which distribution shows the most variability?



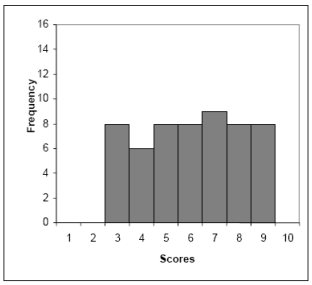
I



II



III



IV

Please check the answer that describes your level of confidence in the answer you picked.

	I feel clueless about the answer.	I think this may be the right answer.	I feel pretty good about the answer.	I am completely sure it is right.
Confidence Level				

15. Which is true of a t-distribution?

a) It is used for small samples
b) It is used when the population standard deviation is not known
c) It has the same basic shape as a normal distribution but has less area in the tails
d) a & b are both true
e) a, b & c are all true

Please check the answer that describes your level of confidence in the answer you picked.

	I feel clueless about the answer.	I think this may be the right answer.	I feel pretty good about the answer.	I am completely sure it is right.
Confidence Level				

18. A researcher performs a t-test to test the following hypotheses:

$H : \mu \leq \mu$
$H : \mu > \mu$

He rejects the null hypothesis and reports a p-value of 0.10. Which of the following must be correct?

a) The test statistic fell within the rejection region at the $\alpha = 0.05$ significance level
b) The power of the test statistic used was 90%
c) Assuming H0 is true, there is a 10% possibility that the observed value is due to chance
d) The probability that the null hypothesis is not true is 0.10