Flipping the Classroom for Introduction to Probability and Statistics for Engineers

Philomena Krosmico

West Virginia University, philomena.krosmico@mail.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Part of the Engineering Education Commons, and the Industrial Engineering Commons

Recommended Citation
Krosmico, Philomena, "Flipping the Classroom for Introduction to Probability and Statistics for Engineers" (2021). Graduate Theses, Dissertations, and Problem Reports. 8277.
https://researchrepository.wvu.edu/etd/8277

This Problem/Project Report is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Problem/Project Report in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Problem/Project Report has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.
Flipping the Classroom for Introduction to Probability and Statistics for Engineers

Philomena Krosmico

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Part of the Engineering Education Commons, and the Industrial Engineering Commons
Flipping the Classroom for
Introduction to Probability and Statistics for Engineers

Philomena Krosmico

Problem Report submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources

in partial fulfillment of the requirements for the degree of

Master of Science in Industrial Engineering

Committee:

Kenneth Currie, Ph.D., Chair
Majid Jaridi, Ph.D.
David Wyrick, Ph.D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia
2021

Keywords: flipped classroom, inverted classroom, engineering statistics learning

Copyright 2021 Philomena Krosmico
ABSTRACT

Flipping the Classroom for Introduction to Probability and Statistics for Engineers

Philomena Krosmico

Introduction to Probability and Statistics for Engineers, IENG 213, is a foundational course in the Industrial Engineering curriculum at West Virginia University (WVU). The challenge has been finding the best teaching method to instill concept learning. A “flipped classroom” teaching style has been gaining momentum throughout higher education and has had proven success in STEM fields. At WVU, beginning with the 2016 Fall semester, the teaching model for the class used a flipped classroom style for one of the instructors. Data on statistical concept learning, using the University of Oklahoma validated statistics concept inventory instrument (Allen K., 2006), was collected for five semesters for a "flipped" classroom and two semesters for a traditional lecture classroom. This difference in class type, coupled with the concept inventory results has provided an opportunity to compare the pedagogical effect on student concept learning.

In addition, this research has expanded the current body of research by including two math factors – (1) Initial math level of engineering students at the start of their academic career and, (2) an aggregate math grade through Calculus 2, which is the co-requisite for IENG 213. This study infers the influence of math preparedness and the effectiveness of the type of instructional methodology on student concept learning. This analysis shows that depending on initial math level and math grade, certain students on average improve concept learning in a flipped classroom. It shows that for the highest math levels or math grades mean concept learning, teaching style has no significant effect. It also shows for the lowest math levels, the mean concept teaching style has no significant effect.

However, students in the moderate math level and low and moderate math grades do show improved concept learning in the flipped classroom. In all the analysis in this study, the “Flipped” classroom was either neutral or greater in post-test concept scores when compared to the Traditional classroom. “Flipped” was never inferior to the “Traditional” classroom in statistical concept learning.
# TABLE OF CONTENTS

ABSTRACT.................................................................................................................................................. ii
LIST OF TABLES......................................................................................................................................... iv
TABLE OF FIGURES .................................................................................................................................... v
INTRODUCTION........................................................................................................................................... 1
Concept Inventory Development and Evolution (the University of Oklahoma research) .......... 2
SCI Assessment Results for Reliability and Validity............................................................................... 5
FtC Studies and Impact on Student Learning......................................................................................... 6
RESEARCH PURPOSE AND QUESTIONS................................................................................................. 7
RESEARCH METHODOLOGY AND RESULTS ...................................................................................... 8
CONCLUSIONS.......................................................................................................................................... 30
BIBLIOGRAPHY......................................................................................................................................... 33
APPENDIX................................................................................................................................................ 34
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exam Question Mean Comparisons between Traditional &amp; FtC Classes</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Least Square Matrix Results - Model 1</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td># of Observations for 3-Way Combination</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>Least Square Matrix Result 3-Way Interaction-Mod 2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td># Observations/Combination-Model 2</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>Least Squares Means Matrix and Confidence Intervals for 3-Way Interaction Model 1</td>
<td>34</td>
</tr>
<tr>
<td>7</td>
<td>Least Square Means Matrix and Confidence Limits for 3-Way Interaction Model 2</td>
<td>35</td>
</tr>
</tbody>
</table>
TABLE OF FIGURES

Figure 1: Psycho-Educational Origins of Student-Centered Learning Theories ........................................ 2
Figure 2: Paired t-Test Results Comparing Pre and Post Test Results .................................................. 8
Figure 3: Least Square Mean Pre-Test Comparisons
Figure 4: Final Grade IENG 213 Comparison Fall 2018 ........................................................................ 10
Figure 5: FtC vs Traditional Post-Test Concept Mean Comparison .............................................................. 11
Figure 6: Residual Normal Quantile Plot - Model 1 ............................................................................. 14
Figure 7: Residual by Predicted Plot - Model 1 .................................................................................. 14
Figure 8: Clnv Final Residuals vs Class Type .................................................................................. 14
Figure 9: Clnv Final Residual vs Math Level .................................................................................. 15
Figure 10: Clnv Final Residual vs Math Grade .................................................................................. 15
Figure 11: Analysis of Variance Model 1 .................................................................................. 15
Figure 12: Effect Tests and Effect Summary - Model 1 ........................................................................ 16
Figure 13: Class Type Mean Comparison - M1
Figure 14: Math Level Mean Comparison - M1 .............................................................................. 17
Figure 15: Math Level*Math Grade Mean Comparison - Model 1 ........................................................... 18
Figure 16: CType*MLevel*MGrade Means Compare-Mod 1 ................................................................. 19
Figure 17: Clnv Final vs CT*ML*MG Means Plot-Mod 1 ................................................................... 20
Figure 18: Residuals Normal Quantile Plot - Model 2 ........................................................................ 21
Figure 19: Residuals by Predicted Plot - Model 2 .............................................................................. 22
Figure 20: Residual by Class Type-Mod 2 ...................................................................................... 22
Figure 21: Residual by Math Level-Mod 2 ........................................................................................ 22
Figure 22: Residual by Math Grade-Mod 2 ...................................................................................... 23
Figure 23: ANOVA Table - Model 2 .............................................................................................. 23
Figure 24: Effect Tests and Effect Summary - Model 2 ........................................................................ 24
Figure 25: Class Type Mean Compare - M2
Figure 26: Math Level Mean Compare - M2 ................................................................................... 25
Figure 27: Least Square Mean Comparisons Class Type*Math Grade - Model 2 ............................... 26
Figure 28: Least Square Mean Comparisons Math Level*Math Grade - Model 2 .............................. 26
Figure 29: Least Square Mean Comparison 3-Way Interaction-Mod 2
Figure 30: Clnv Final Means vs CT*ML*MG - Mod 2 ........................................................................ 27
Figure 31: Class Type Mod 1
Figure 35: Class Type Mod 2 ........................................................................................................ 28
Figure 32: Math Level Mod 1
Figure 36: Math Level Mod 2 ........................................................................................................ 28
Figure 33: ML*MG Mod 1
Figure 37: ML*MG Mod 2 ........................................................................................................ 28
Figure 34: 3-Way Inter Mod 1
Figure 38: 3-Way Inter Mod 2 ........................................................................................................ 29
INTRODUCTION

Over the past ten to fifteen years, educators have been evolving a teaching style often referred to as “Flipping the Classroom” (FtC) across all levels of education from K-16. This evolution has migrated into more STEM higher education fields including engineering. There are numerous articles, conference proceedings, and research supporting the FtC teaching style. The recent online push for education during the pandemic has also escalated the opportunity to integrate learnings from FtC into online classes.

FtC is the process where the lecture is previewed by the students before the start of each class and the homework, or additional enhancement activities are performed inside the classroom. This allows instructors to facilitate and assist students while observing their work. The pre-work allows students to become familiar with the materials and then use class time to focus on issues of concern or more complex problems, under the instructor’s guidance (Karabulut-Ilgu, Cherrez, & Jahren, 2018) (Bishop, 2013). Lecture outside of class can be video lectures, lecture notes, exercises, and/or assignments. Many textbook companies have developed excellent materials specifically designed for use in this type of teaching style (Clark & Kaw, 2019). Some classrooms establish groups or teams inside the class to solve problems or work on projects (Olsen, 2014). The intent is to elevate motivation and engage students in their own learning.

Kahn Academy, founded by Sal Kahn, helped escalate this type of learning style and transform education (Bishop, 2013). From 2009 through today, this teaching style has been on a rapid pace of implementation and analytics.

In many of the studies, feedback surveys were given to students to determine if they preferred this teaching style for their learning (Olsen, 2014) (Gunduz & Akkoyunlu, 2019). Results were often mixed as expected (Vidic, Clark, & Claypool, 2015). Some students preferred it, and others did not (Olsen, 2014). There were concerns over the amount of work required with the pre-work before the class and the classwork in the class. There were also concerns that when the lecture was via video, there was not an instructor there to immediately, answer questions (Gunduz & Akkoyunlu, 2019) (Hotle & Garrow, 2015). This feedback mimics the feedback received for the seven semesters of IENG 213 FtC at WVU. Students who are highly motivated and engaged, come to class prepared and ready to actively participate (Fedesco & Troy, 2016). These types of students are generally positive about the experience and their own learning. In addition, when past students return to the instructor to get copies of the videos to help them with other classes, it reinforces the process. This happened multiple times at WVU over the past four years.

Many of the studies do not have a quantitative conclusion citing FtC as improving learning of the subject matter at hand (Fedesco & Troy, 2016), but rather citing qualitative improvements. However, there are other peripheral benefits to the FtC style – it motivates students to come to class prepared, every class. This reinforces positive professional standards for future work environments, which will require them to prepare and participate actively. In addition, leadership, collaboration, and teamwork are more important today to create the efficiencies and flow needed to execute effective business plans. FtC interactions allow students to function in that type of environment.
The research articles on FtC generally ignore the importance or type of “in-classroom” work, and the primary focus is usually on the development of video lectures (Bishop, 2013). The "in-classroom work" is equally an important focus for the student’s learning. ASEE is recommending (Bishop, 2013) more research be done focused on the “in-classroom” activities. While all the cases Bishop reviewed had both the video lecture and classroom work, details behind this classwork were lacking. They also recommended a check on problem style questions throughout the semester, as well as concept inventory questions (Bishop, 2013). The classroom is the place where professors/instructors get the opportunity to build that relationship and close the loop with the students. How people learn is certainly complex and different for everyone. It often takes explaining it or asking it in several different ways, until you can find the option that makes the "light bulb" go on.

Figure 1 below depicts the complexity of learning and shows the learnings and progression over time. Learning has been a focus of educators for decades. Figure 1 (Bishop, 2013) shows the beginning in the mid-1930s progressing to more experiential learning in the mid-1980s, collaborative learning beginning in the early 1990s, and progressing into the 2000s. As an instructor in a FtC environment, you begin to understand the differences of each students’ learning and what will have the best chance for success (Fedesco & Troy, 2016). The FtC atmosphere can be very engaging and non-threatening, creating an informal environment for students to learn, ask questions and work together with their instructor, peers, and teaching assistant.

Concept Inventory Development and Evolution (the University of Oklahoma research)

By definition, a concept inventory is a criterion-referenced test designed to help determine whether a student has an accurate working knowledge of a specific set of concepts. Concept inventories tend to be a multiple-choice format, for ease of administering, especially in large classes, and for interpretation of results. “The concept inventory movement was spurred by the development and successful implementation of the Force
Concept Inventory (FCI) (Allen K., 2006) (Stone A., 2006). The FCI was developed as a pre/post-test to identify student misconceptions about Newtonian force when entering a physics course and check for gains upon completing the course. After many rounds of testing, it was discovered that students gain the most conceptual knowledge through interactive engagement courses, as opposed to traditional lectures" (Hake, 1998) (Allen K., 2006). The development and validation for this physics FCI took more than 16 years to complete.

The success of this FCI began a movement to develop concept inventory instruments in other science, mathematical, and engineering fields of study (Allen K. S., 2004). "In light of recent Accreditation Board for Engineering and Technology (ABET) standards which focus on outcomes rather than simply fulfilling seat time requirements, many engineering fields have developed concept inventories, such as Thermodynamics, Statics, and Heat Transfer" (Evans, 2003). After several rounds of development and testing, it was eventually concluded that students gain the most conceptual knowledge in interactive engagement courses versus traditional lectures (Bishop, 2013). This also prompted educators to develop teaching techniques to be more interactive, collaborative, and engaging.


The pilot version was constructed by first identifying topics to include using a modified Delphi approach. 23 professors were surveyed to help determine topics that should be included in the concept inventory instrument (Stone A., 2006). In addition, a review of the Advanced Placement (AP) Statistics Course Description identified four general topic areas broken into 20 sub-topics (Stone A., 2006). The two lists were compared, and both the faculty list and AP Statistics list were very similar (Allen K., 2006). This list aligns well with the format of the WVU IENG 213 class.

Development of the initial set of questions, in multiple-choice format, began by researching multiple statistics textbooks, other educational literature for examples related to the topics identified in the list above, and the researchers' personal experience (Allen K., 2006). The refinement of the questions went through several rounds of analysis. First, they assessed test takers' choice selection. If there were answers that were consistently not chosen, those answers were removed or changed. (Allen K., 2006) Second, focus groups with students were held to understand and gain insight as to why they selected certain answers. When changing or creating new questions, the researchers were careful in the development not to create questions that were definitions, or problem-solving. (Allen K., 2006) They purposefully created questions that tested concepts, not recall or computations. (Allen K., 2006)

Careful thought and research were conducted into how the questions should be worded as not to give an advantage to students with good test-taking skills. There were seven criteria cited by Gibb in his 1964 dissertation, listed below, that OU researchers used when reviewing the wording of the questions and solutions (Gibb, 1964) (Allen K., 2006) (Stone A., 2006):
1. Phrase-Repeat: Correct answer contains a key sound, word, or phrase contained in the question’s stem. The stem is the part of the question that presents the issue that the question is asking.
2. Absurd Relationship: Distracters are unrelated to the stem.
3. Categorical Exclusive: Distracters contain words such as “all” or “every.”
4. Precise: Correct answer is more precise, clear, or qualified than the distracters.
5. Length: Correct answer is longer than the distracters.
6. Grammar: Distracters do not match the verb tense of the stem, or there is not a match between articles (“a”, “an”, “the”).
7. Give-Away: Correct answer is given away by another item in the test.

Determining validity is based on collective evidence from multiple sources. “Reliability is a necessary component of test validity, but it is not sufficient evidence of validity. Messick (1989) discusses three components of construct validity which should be addressed when assessing validity claims: a substantive component, a structural component, and an external component.” (Stone A., 2006)

“The most common measure of reliability is Cronbach's coefficient alpha. The coefficient of alpha (Cronbach alpha) and discrimination was used to measure question validity. Alpha-if-deleted was assessed. If students with lower test scores perform better than students with higher test scores on a question, the question was analyzed and either deleted or changed.” (Stone A., 2006)

Discrimination is the test’s ability to produce a wide range of scores. When assessing the differences between student subjects, it is preferred to have a wider range of scores. As stated above, the SCI evolved over multiple iterations where questions were thoroughly vetted to include only concept questions that were worded not to lead good test takers or good guessers to the solution. Solutions were also thoroughly analyzed for each question, especially through focus groups with students to understand “why” they picked certain answers. Experts were involved in developing the subject matter and questions, as well as the AP Statistics Course Description (Stone A., 2006).

External validity was also assessed. That is, how well the SCI correlates to other components. In this case, other components could be “other tests” or “course grades”. “The ‘other test’ in this case would be the overall course percentage grade, which is correlated with the SCI Pre Test, SCI Post-Test, SCI Gain (Post minus Pre), and SCI Normalized Gain (Gain as a percentage of the maximum possible Gain). Based on the results of the full-information Maximum-Likelihood Factor Analysis, the instrument is also divided into four sub-tests, Descriptive Statistics, Probability, Inferential and Graphical Analysis. The concurrent validity of the sub-tests is assessed by correlating the overall course grade with the score on each of the sub-tests.” (Allen K. S., 2004) Initially, this was not conclusive. However, after several iterations, in the Fall of 2003, four engineering statistics courses were assessed, and one math statistics was assessed. The results showed that two of the engineering pre-tests correlated with the final grade and all four of the engineering post-tests correlated with the final grade. However, none of the math final grades correlated with either the post or the pre-tests. Since the SCI was designed for Engineering Statistics, and engineering is more applied than theoretical, the results were acceptable. Of the engineering classes assessed, 3 of the 4 were external universities and 1 internal to the University of Oklahoma (Allen K. S., 2004).
SCI Assessment Results for Reliability and Validity

After assessing the first four semesters of the SCI, and refining the questions and answers, the results indicate that the instrument improved in terms of validity, reliability, and discriminatory power (Allen K. S., 2004). Multiple iterations produced the concept inventory used today. Over 4 years (7 semesters), 1500+ students from Oklahoma University and 7 External colleges/universities have used the concept inventories (Allen K., 2006). Every question has had some modification for improvement in reliability, validity, or discriminatory power. Ongoing changes made were measured to determine if the SCI instrument is considered reliable and valid in determining introductory statistics concept knowledge. A variety of methods were used for this assessment.

As stated above, a common measurement for reliability is Cronbach’s coefficient of alpha. The expected range for the success of this model was between 0.60 and 0.80. The overall instrument result is at 0.70 (Stone A., 2006). All individual items on the latest version of the instrument were tested for “alpha-if-item-deleted” and are all within the acceptable range. Individual items were also measured to determine if they have discriminatory power. Items analysis showed all were 0.90 or greater (Stone A., 2006). Another method evaluated to prove reliability is called item response theory (IRT). IRT methods model the probabilities of a correct response using nonlinear models. Instead of using the overall test score for each examinee, the IRT analysis measures the pattern of responses to each test item for each subject. Ultimately, it models the probability that the subject will answer the specific item correctly. The overall result is 0.787. (Stone A., 2006) This result is based on the results of all assessments given of all the administrations of the SCI, not just a single administration. The author, Andrea Stone, referenced Nunnally’s book from 1967, Psychometric Theory, and concluded this result is considered adequate to conclude reliability. (Stone A., 2006)

In addition to the reliability measurements above, validity is also assessed. All questions have been thoroughly evaluated and modified to measure concepts, not definitions or computations. Item analysis included response distributions, evaluation of answer distractions, and focus groups across all subjects both internal and external assessment takers were included to refine and improve the items used on the inventory. All items on the SCI are now considered concept questions and have correlated with the final post-test scores.

Factor loadings were also assessed. A factor loading is a measure of a given attribute. In this case, the attribute would be a conceptual understanding of introductory statistics. All factor loadings were positive indicating the total score is a reasonable estimate of introductory statistics. (Stone A., 2006)

According to Stone she writes, "it has been demonstrated to be a reasonably reliable instrument for research use. The SCI should be used in classroom settings as a post-test and optionally as a pre-test to evaluate instructional methods. Baseline data is available that can be used as a benchmark for comparison." (Stone A., 2006)

FtC Studies and Impact on Student Learning

In 2019, two researchers from The University of Hong Kong reviewed the impact of the
FtC on student achievement in engineering education. They reviewed ten years of research and focused on studies that included numerical results. Their analysis compared 29 international peer-reviewed research papers comparing traditional and FtC concerning student achievement. The results indicated that there was a significant effect in favor of the FtC methodology (Lo & Hew, 2019).

While research has supported FtC style of teaching in many fields of study, there has only been one paper in the field of Engineering Statistics, however not specifically using a validated and reliable student learning concept inventory. The University of San Diego flipped Engineering Probability and Statistics and used homework, quizzes and exam grades to compare but did not prove a significant difference in teaching styles. (Olsen, 2014)

The overall benefit to student learning is the result in the majority of studies, albeit by different methods of quantitative and qualitative measurement. Most studies that have a quantitative analysis of results use similar test questions or exams to measure learning. Many studies use qualitative results to assess learning. Finding the best consistent measurement tool or process to assess learning is one of the gaps in current research. Another gap in the research is identifying specifically which “type” of students gain the most from this teaching style and quantify specifically what about the teaching style makes it most effective.

**FtC Studies Using the University of Oklahoma Concept Inventory**

The University of Pittsburgh researchers (Vidic, Clark, & Claypool, 2015) found that Engineering Statistics did not prove a significant difference in learning related to the difference in teaching styles (Vidic, Clark, & Claypool, 2015). The research compared the FtC methodology to the traditional lecture-style and measured results using the Oklahoma University concept inventory, assessing common questions between traditional lecture classes and FtC classes. The University of Pittsburgh had several instructors involved. For the FtC classrooms, they flipped four weeks of a class (4 chapters), and then the remaining lectures in the class were traditional "not flip" lectures. In addition to the University of Oklahoma SCI, two test questions were common across both “flipped” and traditional lecture classes. Instructors were different between the “flipped” and “traditional” and, also different within “flipped” and “traditional”. They concluded there was no evidence to support a significant difference between lecture styles. (Vidic, Clark, & Claypool, 2015)

The University of Pittsburgh’s execution of FtC is different than the study analyzed in this paper. The focus of the WVU research study used FtC for the full semester (16 weeks) in one section and traditional lecture for the full semester in another section. The FtC instructor was consistent, and the traditional lecture instructor was consistent. The structure of the classes and consistent instructors might be a partial determinant in these conclusions.

Based on the reliability results above, extensive validation of individual items through multiple means, factor loadings, and discriminatory indices, I have chosen to use this SCI to evaluate the potential difference in FtC instructional style compared to the traditional lecture instructional style for this research.
RESEARCH PURPOSE AND QUESTIONS

At WVU, the FtC format was established for Introductory Probability and Statistics for Engineers (IENG 213) in the Fall of 2016. Videos, lecture notes, and quizzes/homework were created and have continued to evolve and improve. Key concepts of the video lecture are reviewed at the start of the class to answer any questions and emphasize key learning points. Recent literature commented that this in-class review of “flipped” content at the beginning of the class is a key factor in learning (Lo & Hew, 2019). In the IENG 213 class, students work in teams of three or four students. The instructor and teaching assistant work with teams to understand the gaps in their learning. Semester to semester this has allowed the material to evolve to help close the learning gaps. I have personally observed the qualitative benefits of this teaching style for both the student and the instructor that is highlighted throughout the research. Finding the best method of student concept learning is important. One of the areas assessed in this research is the “type” of student that gains the most from this teaching style. Math level and math grade of a student will be assessed to determine if they are factors in improved concept learning through the FtC model for the specific class, Introduction to Probability and Statistics for Engineers.

This research addresses the following questions for the Introduction to Probability and Statistics for Engineers class (IENG 213):

1. Is the FtC model effective for student learning in Introduction to Probability and Statistics for Engineers?
2. Does the FtC model for Introduction to Probability and Statistics for Engineers result in improved learning compared to the traditional, "not flip" format?
3. Is there a significant improvement in learning with a FtC classroom format, for students at specific math grades or students who started in certain levels of math?

The hypotheses assessed for Question 1, 2, and 3, respectively are:

Question 1:
H0: There is no significant difference between concept inventory initial scores (pre-test) and final scores (post-test) for students in the FtC.
H1: Concept learning final scores (post-test) are statistically different than initial scores (pre-test) for a student in the FtC.

Question 2:
H0: There is no difference in concept learning or exam questions between the Probability and Statistics for Engineers FtC class and the traditional lecture class.
H1: The results measured in final concept inventory comparisons or exam questions are statistically different for the FtC class versus the traditional lecture class.

Question 3:
H0: The Concept Inventory test scores at the end of the class were the same between FtC and traditional classes, given the same math grades and starting math levels.
H1: The Concept Inventory test scores were statistically different for FtC classes given the same math grades and starting math levels.
RESEARCH METHODOLOGY AND RESULTS

While some previous research has partially addressed Questions 1 and 2, Question 3 is new research with respect to the FtC topic. This was measured by establishing math levels and math grade scores. These factors coupled with class type – FtC, or traditional lecture classroom formats – are compared to see if there is a difference in learning based on mean post-test (final) concept inventory scores.

An IRB was written and approved to use the Concept Inventory tool. Students received a letter stating the use of the concept inventory tool was for research. Their identities were hidden and if they chose not to participate, they did not have to. There was no incentive given to those choosing to take the concept inventory. Simultaneous classes using the FtC, and traditional lecture models, were taught in the semester once approval was received to use the University of Oklahoma SCI. At the end of the Fall 2017 semester, a final concept inventory (post-test) was given to both sets of students. This was the beginning of the process to assess concept learning using the University of Oklahoma instrument.

The process involves an initial concept inventory assessment (pre-test) on the first day of class and again at the end of the semester (post-test) during the last week of class. The post-test is considered the “final” concept inventory. This continued each semester from the Spring of 2018 through the Fall of 2019. In all cases, fewer students chose to take the post-test than the pre-test. Feedback from students indicated that by the end of the semester, they were “test weary” from taking tests for all their other classes. Since the post-test is not required, and they did not get credit for taking it, they chose not to take it. One note, the pre-test was given at the beginning of the semester to the Spring class of 2020 students but was not given as a post-test due to the interruption of classes and fully online learning due to the pandemic. Due to maintaining control of the document, it was determined not to create an online “test” copy of the instrument.

**H0: There is no significant difference between concept inventory initial scores (pre-test) and final scores (post-test) for students in the flipped classroom**

To address Question 1 (Hypothesis 1), pre- and post-tests were given at the beginning and end of the semester respectively for all FtC classes from Spring 2018 through Fall 2019. A paired t-comparison per student for students who completed both an initial and final concept inventory assessment for each semester was analyzed. 79 students completed both the pre-test and post-test and the paired data t-test comparison results (Figure 2), show a significant difference in learning, with a p-value of 0.0124.

![Figure 2: Paired t-Test Results Comparing Pre and Post Test Results](image-url)
**H₀: There is no difference in concept learning or exam questions between the Probability and Statistics for Engineers FtC class and the traditional lecture class**

To address question 2, two semesters (Fall of 2017 and Fall of 2018) were set up to offer both FtC and traditional lecture sections so that the mean test results could be compared. The FtC was offered Tuesday and Thursday, 8:00 – 9:15 am and the traditional lecture was offered Monday, Wednesday, and Friday, 8:00 – 8:50 am. The instructors were different but worked together to cover the same material including the use of 21 of the exact same in-class exam questions.

As stated above, pre-tests were given for all semesters of the FtC from Spring 2018 through Spring 2020. The initial inventory for the Fall 2018 traditional lecture classroom scored 46.8% lower than the mean of all the other initial inventories. When the data was analyzed, it showed that 76% of the students (32 out of 42) only finished at most 2/3 of the pre-test, indicating not enough time was given to complete the assessment. In all the other semesters, only 3.5% (13 out of 368) left more than one question unanswered. It was determined that the initial inventory data for the traditional lecture Fall 2018 classroom was not representative and therefore not valid. The results for the pre-test for that specific classroom are not considered in this analysis.

In order to compare post-test concept results, it was important to determine if there was a significant difference in concept understanding at the beginning of the semester between any sections of the class who have taken the concept inventory pre-test. Pre-test concept scores were analyzed for all the FtC semesters from Spring 2018 through Spring 2020, to determine if any of the semesters started with a higher mean concept score at the beginning of the semester. A JMP test measuring the differences in means compared valid data for all semesters and the results indicate that there was no significant difference in the pre-test scores (Figure 3). In addition, final IENG 213 class grades were compared for the Fall 2018 Semester between the FtC and traditional lecture classes. The mean Grade Point Average (GPA) for the FtC was 3.1 and the mean GPA for the traditional lecture class was 3.3. When compared statistically using a difference between mean test (Figure 4), there was no significant difference detected with a p-value = 0.34. Based on both results, it is concluded that the average initial Fall 2018 traditional class concept knowledge was not statistically lower than the average initial concept Fall 2018 FtC knowledge, and the post-test results can be compared to determine if there is a significant difference in post-test mean results.
To evaluate question 2, post-test concept inventories were given to both Fall 2017 and Fall 2018 FtC and traditional lecture classes at the end of the semester (post-test). This is one of the tests used for assessing the hypothesis for question 2. If the results measured in the final concept inventories comparisons or exam questions are significantly different for the FtC versus the Traditional class for the Probability and Statistics for Engineers, we will reject the null hypothesis based on sufficient information to support the alternative hypothesis. 60 students participated from the Traditional classroom (Type 1) and 80 students participated from the FtC (Type 2).

A JMP t-test means comparison shows that the mean post-score is statistically different between the classes, favoring FtC since the mean FtC was significantly greater than the traditional. The p-value is 0.0014 (Figure 5).

Interpreting the JMP output, there is a 95% confidence level that the mean post-test score for the traditional classroom ranged from 37.2 to 42.7, while the mean post-test score for the FtC ranged from 43.6 to 48.4. There is also a 95% confidence level that the mean difference in post-test scores between the traditional lecture format and FtC ranges from 2.4 to 9.7 points, in favor of the FtC style.
In addition, during the fall of 2018, exam questions were coordinated between both instructors, and the same grader graded both class sections of the questions. Students in the Tuesday/Thursday section (FtC) had 75 minutes to take tests versus the lecture section had 50 minutes. Note: Additional questions were on the FtC tests to compensate for the additional time. Of the twenty-one topics assessed, fourteen came back no difference, six were in favor of the FtC class and one was in favor of the traditional lecture class. The significance results are summarized in Table 1.
Table 1: Exam Question Mean Comparisons between Traditional & FtC Classes

Yellow highlights favor FtC class; Blue highlights favor Traditional class

<table>
<thead>
<tr>
<th></th>
<th>Traditional Lec Class</th>
<th>FtC Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Score</td>
<td>n</td>
</tr>
<tr>
<td>Confidence Interval Variance</td>
<td>84.636</td>
<td>55</td>
</tr>
<tr>
<td>Hypothesis Test - Proportion</td>
<td>83.727</td>
<td>55</td>
</tr>
<tr>
<td>Normal Probability Problem</td>
<td>84.729</td>
<td>58</td>
</tr>
<tr>
<td>Discrete Probability Problem</td>
<td>50.677</td>
<td>57</td>
</tr>
<tr>
<td>Cumulative Probability</td>
<td>76.608</td>
<td>57</td>
</tr>
<tr>
<td>Joint Probability</td>
<td>69.424</td>
<td>57</td>
</tr>
<tr>
<td>k value (area) - Z table</td>
<td>90.887</td>
<td>58</td>
</tr>
</tbody>
</table>

H₀: The Concept Inventory test scores at the end of the class were the same between FtC and traditional classes, given the same math grades and starting math levels

Question 3 asks – **Is there a significant difference in learning with a FtC classroom format, for students at specific math grades or students who started in certain levels of math?**

To answer this question additional data collection relating to math performance was necessary. Data on initial math level and math grades were collected for students taking Introduction to Probability and Statistics between the 2017 fall semester and the 2019 fall semester.

Math Level for the purpose of this study is the level at which an engineering student is placed into upon entering the Statler College of Engineering. When a student is accepted into the Statler College of Engineering and placed into a math level, it is based on ALEKS, ACT or SAT math test results. Initial models developed used five levels:

- Pre-Calculus Level 1 – Began at MATH 126/128 (Algebra/Trigonometry)
- Pre-Calculus Level 2 – Began at MATH 153/154 (Pre-Calculus)
- Calculus Level 3 – Began at MATH 155 (Calculus 1)
- Calculus Level 4 – AP tested out of MATH 155 (Calculus 1)
- Calculus Level 5 – AP tested out of MATH 155 and 156 (Calculus 1 & 2)

The second factor was Math Grade score. Math Grade score was based on grades for pre-calculus through Math 156. Pre-Calc, Calc 1, and Calc 2 were the only grades considered. Calc 2 was chosen as the cutoff grade since this is currently a co-requisite for IENG 213. If the average was at the mid-point or below, the lower score was given. If it was higher than the mid-point, the higher score was given. For example, if a student had taken Calc 1 and Calc 2 and got an A and B, they were scored as a B. If they had taken Pre-Calc, Calc 1, and Calc 2, and got two A’s and one B, they were scored as an "A." Math scores were assigned as follows:

- 4 for an A
- 3 for a B
- 2 for a C
- 1 for a D
In addition to Math Level and Math Grade, students were coded for Class Type (1 for traditional lecture or 2 for FtC).

When collecting data on Math Level and Math Grade, it was discovered that some Biomedical Engineering students had taken STAT 215 prior to IENG 213. Students can elect to take both as independent electives. Since these two classes are very similar, it was decided to eliminate students that took STAT 215 before taking IENG 213. Since they already had a very similar course, it could bias the concept inventory scores when comparing to students that had not taken STAT 215 before IENG 213. This resulted in removing three students from the study (they were removed from all analyses for all hypotheses evaluated). Students taking IENG 213 and STAT 215 simultaneously stayed in the data pool.

There were 209 students across all the semesters that had final concept inventory scores. When evaluating the initial levels established, there were a small number of students in several categories, and the decision was made to reduce the number of levels for both Math Level and Math Grade.

For Math Level, there were only three students in Level 1 pre-calculus, so it was reduced to only one level associated with pre-calculus. There were two students who had AP credit for both Calc 1 & 2, it was decided to reduce to one level for AP credit. The three math levels are defined as follows:

- 4 for AP for Calculus 1 or Calculus 1/Calculus 2
- 3 for Calculus 1
- 2 for Pre-Calculus

Three students had grades categorized as D's. Each data record was revisited. For students to progress forward to take the third level of Calculus they must achieve a C-math grade. Since the number was small and their grades were between C's and D's, it was determined to include them in the C category. The three math grade levels are defined as follows:

- 4 for an A
- 3 for a B
- 2 for a C or D

Once the data was cleansed, the data records were re-assembled without student names or numbers. A Least Squares model with Effect Leverage emphasis was run on JMP software with a full factorial format to capture the interactions of the main effects. The response variable was CI Final (concept inventory final post-test score), and the main effects were Class Type, Math Level, and Math Grade.

Residuals were analyzed for normality, constant variance, and independence. The Residual Normal Quantile Plot (Figure 6) shows that some residuals touch and slightly cross, but the majority are within the curves. The residual by predicted plot (Figure 7) shows a slight variation in the variance for CI Final Result of approximately 30 or less. Most of the results are over 30 and it was determined not to transform the model for this slight variation.
In addition, the residuals versus main effects (Figures 8-10) show slightly less variance in Type 1 (traditional lecture class) on the high end and for Math Level 4 on the low end. These variations are small and considered insignificant.
The model was determined to be a good model to use. The ANOVA table (Figure 11) results below show significance, at a p-value < 0.0001.
The significant effects are shown in Figure 12, Class Type, Math Level and the two-way interaction, Math Level*Math Grade are significant < 0.05 p-value. The three-way interaction (Class Type*Math Level*Math Grade) p-value is relatively small at 0.0968 and will be evaluated further to determine if any of the three-way combinations can be determined to be significant.

![Effect Tests and Effect Summary - Model 1](image)

Each significant effect was analyzed to understand the potential differences in mean concept learning. For Class Type 1 (traditional lecture), the least square mean is 38.7 and for the FtC it is 45.3 as shown in Figure 13. It shows there is a significant difference between Class Types. Based on this model, we are 95% confident that the mean difference in post-test scores will be between 1.5 and 11.7 points, in favor of FtC.
Using the same methodology for Math Levels, the results in Figure 14 show there is no significant difference between Level 3 and 4, but there is a significant difference between both Level 3 and 4, and Level 2. The least-square mean for Level 3 is 44.8, Level 4 is 44.5 and Level 2 is 36.7.

We are 95% confident that the mean difference between an initial math level of Pre-Calc and Calc 1 is between 2.8 and 13.4 points. We are 95% confident that the mean difference between an initial math level of Pre-Calc and AP is between 0.5 and 15.1 points.

For Math Level*Math Grade the Least Square Means are shown in Figure 15. Levels not connected by the same letter are significantly different.
Figure 15 can be interpreted as follows:

- For Math Level 3 and 4 (Calc 1 and AP Credit for Calc), there was no significant difference in means when students had a Math Grade of B for Calc 1 or a grade of A, B, or C when starting with AP credit for Calc 1/2.
- There were, however, significant differences between means when students began in Calc 1 with a Math Grade of A, with students starting at Math Level AP with a Math Grade of B.
- There was a significant difference in means for students beginning in Calc 1 with a Math Grade of A or B, or started at AP with an A grade, with students beginning in Calc 1 with a Math Grade of C or students beginning in Pre-Calc with a Math Grade of B.
- There was no significant difference in means for students beginning in Pre-Calc, and students beginning in Calc 1 with a Math Grade of C or AP level with a grade of B or C.
- There was also no significant difference in means for students beginning in Pre-Calc with a Math Grade of A or C, and AP students with a grade of B or C or started in Calc 1 with a Math Grade of B or C.

Three of the confidence levels interpretations, irrespective of Class Type, are as follows:

95% Confident of students starting with a Math Level Calc 1 and a Math Grade of A compared to starting with a Math Level AP and a Math Grade of B the mean difference in post-test scores ranges from 0.5 to 20.0 points, in favor of Calc 1 with a grade of A.

95% Confident of starting with a Math Level AP and a Math Grade of A compared to starting with a Math Level Pre-Calc and a Math Grade of A the mean difference in post-test scores ranges from 3.0 to 29.9 points, in favor of AP Credit with a grade of A.

95% Confident of starting with a Math Level Calc 1 and a Math Grade of B compared to starting with a Math Level Pre-Calc and a Math Grade of B the mean difference post-test scores range from 2.8 to 16.5 points, in favor of Calc 1 with a grade of B.

The three-way interaction, Class Type * Math Level * Math Grade, least square mean comparisons results are shown in Figures 16 & 17. The three-way interaction is in order – class type, math level, math grade. Figure 16 shows the least-squares means and shows letters next to each combination. If levels do not have the same letters connecting them,
they are significantly different. For example, a FtC student who started in Calc 1 with a math grade of B (2,3,3) is statistically different than a FtC student who started in Calc 1 with a math grade of C (2,3,2) because they have different letters in their rows. The full least-square matrix comparison with upper and lower confidence limits at 95% is in the Appendix (Table 6). Figure 17 shows the Least Squares Means Plots for the 3-Way Interaction vs Concept Post-Test Least Square Mean Results.

<table>
<thead>
<tr>
<th>Level</th>
<th>Least Sq Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,4,4</td>
<td>56.111667</td>
</tr>
<tr>
<td>2,3,4</td>
<td>51.667143</td>
</tr>
<tr>
<td>2,4,2</td>
<td>51.111111</td>
</tr>
<tr>
<td>1,3,4</td>
<td>50.556111</td>
</tr>
<tr>
<td>2,3,3</td>
<td>50.114943</td>
</tr>
<tr>
<td>1,2,2</td>
<td>45.238095</td>
</tr>
<tr>
<td>1,4,4</td>
<td>44.999444</td>
</tr>
<tr>
<td>2,3,2</td>
<td>41.720645</td>
</tr>
<tr>
<td>2,2,4</td>
<td>41.539487</td>
</tr>
<tr>
<td>1,4,3</td>
<td>40.834167</td>
</tr>
<tr>
<td>2,4,3</td>
<td>40.832500</td>
</tr>
<tr>
<td>1,3,3</td>
<td>39.998788</td>
</tr>
<tr>
<td>2,2,3</td>
<td>38.778556</td>
</tr>
<tr>
<td>2,2,2</td>
<td>36.249583</td>
</tr>
<tr>
<td>1,3,2</td>
<td>35.000152</td>
</tr>
<tr>
<td>1,4,2</td>
<td>33.333333</td>
</tr>
<tr>
<td>1,2,3</td>
<td>31.999333</td>
</tr>
<tr>
<td>1,2,4</td>
<td>26.670000</td>
</tr>
</tbody>
</table>

Levels not connected by same letter are significantly different.

Figure 16: CType*MLevel*MGrade Means Compare-Mod 1
The results of the Least Square Mean Comparisons when comparing the same Math Level and Math Grade between FtC and Traditional classes are shown in Table 2. The highlighted rows have too few observations in at least one of the means that are insufficient to use. Table 3 shows the number of observations in each category, highlighting the low number of observations.

The results show:

- We are 95% confident that there is a significant difference in post-concept mean scores for a FtC student who started at Calc 1 with a Math Grade of B and a traditional student who started at Calc 1 with a Math Grade of B that ranges from 2.1 to 18.2 points, in favor of the FtC class.
• We are also 95% confident that there is a significant difference in post-concept mean scores for a FtC student who started at Calc 1 with a Math Grade of C and a traditional student who started at Calc 1 with a Math Grade of C that ranges from 0.4 to 13.1 points, in favor of FtC.

While there are multiple other combinations of class type*math level*math grade that are significantly different, these two are the ones that are significant when math level and math grades were held at the same levels, only varying class type.

Since there were concerns with the number of observations in two of the nine comparisons, and for the Main Effect, Math Level, there was no difference between Math Levels 3 (Calc 1) and 4 (AP), a second model (Model 2) was developed combining Math Level 3 and 4 into Math Level 3. Hence, for Model 2, Math Level will now have 2 levels, Class Type and Math Grade will remain the same with 2 levels and 3 levels, respectively.

When Model 2 was run, the residuals were checked as they were in Model 1. Residuals analyzed using a Normal Quantile Plot (Figure 18), Residuals vs Predicted Plot (Figure 19), and the Residuals vs Main Effects (Figures 20-22). Very close to the same results as Model 1.

![Residual Normal Quantile Plot](image)  
*Figure 18: Residuals Normal Quantile Plot - Model 2*
Figure 19: Residuals by Predicted Plot - Model 2

Figure 20: Residual by Class Type-Mod 2

Figure 21: Residual by Math Level-Mod 2
The model was determined to be a good model to use. The ANOVA table (Figure 23) results below show significance, at a p-value < 0.0001.

The effect tests and summary are shown in Figure 24.
The same effects are significant in Model 2 as Model 1, however, Math Level and Math Level*Math Grade have a smaller p-value than Class Type. The three-way interaction is still not significant but is still less than 0.10 at a 0.0941 p-value. The interaction Class Type*Math Grade is less in Model 2 than Model 1, while not significant, is slightly higher than 0.10 at 0.1029 p-value.
The results show a significant difference for both main effects, Class Type and Math Level. For Class Type 1, traditional lecture, the least square mean is 39.4 and for the FtC is 43.9, shown in Figure 25. We are 95% confident that the mean difference in post-test concept scores is between 0.6 and 10.8 points, in favor of FtC.

For Math Level, we are 95% confident that the mean difference in post-test concept scores for an initial Math Level of Calc 1/AP and Pre-Calc is between 2.7 and 12.9 points, in favor of Calc 1/AP (Figure 26).

The new interaction assessed was Class Type*Math Grade. The p-value for this interaction was 0.1029. While not significant, it is small enough to look at the least square mean comparisons. The only combination of Class Type*Math Grade where the grade was constant that showed significance between class types was for a Grade of B (Figure 27). We are 95% confident that there is a significant difference in post-test concept inventory means between students in FtC with a Math Grade of B and the Traditional class with a Math Grade of B with a range of 3.0 to 11.4 points, favoring FtC.
Figure 27: Least Square Mean Comparisons Class Type*Math Grade - Model 2

Levels not connected by same letter are significantly different.

Figure 28 is the Least Square Mean Comparison for Math Level*Math Grade.

Levels not connected by same letter are significantly different.

- For Math Level 3 (Calc 1 & AP) with a Math Grade of A, there is a difference between the post-test concept mean and all other levels and grades.

- For Math Level 3 (Calc 1 & AP) with a Math Grade of B, there is a difference between the post-test mean and Level 3 (Calc 1 & AP) with a Math Grade of C and Level 2 (Pre-Calc) and a Math Grade of B.

For the three-way interaction, Class Type*Math Level*Math Grade, the results are shown in Figures 29 & 30.
The results of the least square mean comparisons when comparing the same Math Level and Math Grade between FtC and Traditional are shown in Table 4. The test is not significant for three of the scenarios compared, significant for two of the six comparisons, and inconclusive for one of the comparisons. The same two scenarios are significant as Model 1, even though Math Level 3 now represents Calc 1/AP. We lost the differentiation at Math Grade B between Calc 1 and AP students. Only one scenario still has a lack of observations. This combination should not be considered valid due to insufficient data and is highlighted in yellow in Tables 4 & 5.

For the significant comparisons, the results show:

- We are 95% confident that there is a significant difference in mean post-test concept scores, for a FtC student who started at Calc 1/AP with a math grade of B and a traditional student who started at Calc 1/AP with a Math Grade of B, ranging from 1.7 to 15.8 points, in favor of FtC.

- We are also 95% confident that there is a significant difference in mean post-test concept scores, for a FtC student who started at Calc 1/AP with a Math Grade of C and a traditional student who started at Calc 1 with a Math Grade of C that ranges from 1.5 to 13.7 points, in favor of FtC.

While there are multiple other combinations of class type*math level*math grade that are
different, these two are the ones that are significant when math level and math grade are held the same. The full three-way matrix comparison for Model 2 is shown in the Appendix (Table 7).

Aside from losing the differentiation of AP students, both models give similar results with some variations in p-values, order of significance and confidence levels. To determine the best model for this study, the parameter powers for the significant effects for both models were compared. Model 1 parameter powers are shown in Figures (31-34) the left-hand column. Model 2 parameter powers are shown in Figures (35-38) the right-hand column. Powers for Class Type and the Three-Way Interaction at $\alpha = .05$ are higher for Model 1 by 20% and 22%, respectively. Powers for Math Level and Math Level*Math Grade at $\alpha = .05$ are higher for Model 2 by 9% and 6% respectively.
<table>
<thead>
<tr>
<th>Figure 37: 3-Way Inter Mod 1</th>
<th>Figure 38: 3-Way Inter Mod 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Power Details</strong></td>
<td><strong>Power Details</strong></td>
</tr>
<tr>
<td>Test Class Type<em>Math Level</em>Math Grade</td>
<td>Test CLASS TYPE<em>MATH LEVEL</em>MATH GRADE</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td></td>
</tr>
<tr>
<td>α</td>
<td>σ</td>
</tr>
<tr>
<td>0.0500</td>
<td>11.51651</td>
</tr>
<tr>
<td>0.1000</td>
<td>11.51651</td>
</tr>
<tr>
<td>0.1500</td>
<td>11.51651</td>
</tr>
<tr>
<td>0.2000</td>
<td>11.51651</td>
</tr>
</tbody>
</table>

| α    | σ     | δ    | Number | Power  |
| 0.0500 | 11.56397 | 1.760291 | 209    | 0.4840 |
| 0.1000 | 11.56397 | 1.760291 | 209    | 0.6111 |
| 0.1500 | 11.56397 | 1.760291 | 209    | 0.6898 |
| 0.2000 | 11.56397 | 1.760291 | 209    | 0.7461 |
CONCLUSIONS

For Question 1, is the flipped classroom model effective for Introduction to Probability and Statistics for Engineers? Yes, we would reject the null hypothesis, based on 79 students in the FtC completing both the pre-test and post-test and the paired data t-test results shown in Figure 2 shows a significant difference in learning, with a p-value of 0.0124. There is sufficient evidence to support and conclude the alternative hypothesis:

\[ H_1: \text{Concept learning final scores (post-test) are statistically different than initial scores (pre-test) for a student in the FtC.} \]

Based on the confidence intervals, we can also conclude that the mean difference in learning (post test score minus pre-test score) is greater than zero for the FtC.

For Question 2, does the FtC model for Introduction to Probability and Statistics for Engineers result in improved learning compared to the traditional lecture format? Yes, based on both the difference between mean t-test comparison analysis of concept learning and the differences using in-class exam problem comparisons.

The t-test for difference between means comparison shows that there is a significant difference in post-score concept learning means between FtC and Traditional classrooms, in favor of FtC. The p-value is 0.0014 (Figure 5). We are 95% confident that the mean post-test score for the traditional classroom ranged from 37.2 to 42.7, while the mean post-test score for the FtC ranged from 43.6 to 48.4. We are 95% confident that the mean difference in post-test scores between traditional and FtC ranges from 2.4 to 9.7 points, in favor of FtC. In addition, of the 21 common exam questions evaluated, 14 showed no difference between mean scores, 7 showed a difference in mean scores, six (6) in favor of FtC, and one (1) in favor of the Traditional class.

Based on both the post-test concept difference in means result and the exam question score comparing difference in means, we would reject the null hypothesis, there is sufficient evidence to support the alternative hypothesis:

\[ H_1: \text{The results measured in final concept inventory comparisons or exam questions are statistically different for the FtC class versus the traditional lecture class.} \]

Based on the confidence intervals, we can also conclude that the mean difference in learning (post test score minus pre-test score) is greater than zero for the FtC versus the traditional lecture format.

For Question 3, is there a significant improvement in learning with a “flipped” classroom format, for students with lower math grades or students who started in lower levels of math?

Two factors involving the initial math level of engineering students and math grade were assessed in conjunction with both class types, FtC and traditional. The response variable is the post-score on the statistics concept inventory. Two models were developed, and both had advantages and disadvantages. The first model had two class types (traditional = Class Type 1 and FtC = Class Type 2), three math levels, and three math grades.
Unfortunately, some of the categories had an unacceptably low number of observations for some of the data comparisons.

To compensate for the low number of observations in some of the deficient categories, math level was reduced from 3 levels to 2 levels, Calc 1/AP and Pre-Calc. The models were very close in their results. However, model 1, while deficient in some of the three-way categories, gave a higher parameter power in both the Class Type mean comparison and the three-way interaction, Class Type*Math Level*Math Grade. The trade-off of number of observations in some deficient categories versus loss of power parameter was not worth the trade. The improvement in power made Model 1 my recommendation as the best model for this research.

Conclusion: Do not reject the null hypothesis for the following three scenarios. There is insufficient evidence to support the alternative hypothesis for the following three specific scenarios.

\[ H_0: \text{The Concept Inventory test scores at the end of the class were the same between FtC and traditional classes, given the same math grades and starting math levels.} \]

If one or both of math level or math grade is high, there is no difference in mean post-scores as outlined below:

- The model showed there is no significant difference in mean post-test scores at the high math level (AP) and high math grade of A, regardless of Class Type.
- The model showed there is no significant difference in mean post-test scores at the high math level (AP) and moderate math grade of B, regardless of Class Type.
- The model showed there is no significant difference in means at the moderate math level (Calc 1) and at the high math grade A, regardless of Class Type.

Conclusion: Reject the null hypothesis and conclude there is sufficient evidence to support the alternative hypothesis for the following two specific scenarios.

\[ H_1: \text{The Concept Inventory test scores were higher for flipped classes given the same math grades and starting math levels.} \]

The model shows the following results at the moderate math level, with either a moderate, or low math grade:

- The model showed there is a significant difference in mean post-test scores at the moderate math level (Calc 1) and moderate math grade of B, in favor of FtC. We are 95% confident that a FtC student who started at Calc 1 with a Math Grade of B is significantly different from a traditional class student who started at Calc 1 with a Math Grade of B that ranges from 2.1 to 18.2 points, in favor of FtC.

- The model showed there is a significant difference in mean post-test scores at the moderate math level (Calc 1) and low math grade of C, in favor of FtC. We are 95% confident that a FtC student who started at Calc 1 with a Math Grade of C is
significantly different than a traditional student who started at Calc 1 with a Math Grade of C that ranges from 0.4 to 13.1 points, in favor of FtC.

Conclusion: Do not reject the null hypothesis, or there is insufficient evidence to support the alternative hypothesis for the following two specific scenarios.

**H₀:** The Concept Inventory test scores at the end of the class were the same between FtC and traditional classes, given the same math grades and starting math levels.

For low Math Level, Pre-Calc, Math Grades B and C:

- The model showed there is no significant difference in mean post-test scores at the low math level (Pre-Calc) and moderate math grade of B, regardless of Class Type.
- The model showed there is no significant difference in mean post-test scores at the low math level (Pre-Calc) and low math grade of C, regardless of Class Type.

**INCONCLUSIVE**

For the following two scenarios, the results are inconclusive due to a low number of observations in these categories.

- High Math Level (AP) with Low Math Grade C
- Low Math Level (Pre-Calc) with High Math Level A

In all cases, as with the rest of this analysis, the FtC classroom was either neutral or results higher in mean post-test concept scores. FtC is never showing up as inferior to the traditional classroom in concept learning. I would conclude this study supports there is a mean difference in post-test concept scores between FtC and Traditional IENG 213 classes. The study also shows that students at moderate math levels and moderate/low math grades on average gain more concept knowledge in the FtC than in Traditional classrooms. Students at high math levels or high math grades, mean post-test concept scores are not significantly different between class types. Students at Pre-Calc math level with moderate and low grades, mean post-test concept scores are not significantly different between class types.

Future research should focus on other "types" of students that FtC may positively impact and what specifically about FtC makes the difference in concept learning. Further assessment of the role of math level in designing the format of the “in class” work and structure of the classroom teams to increase the impact of concept learning. Comparing teams of randomly selected students versus teams where “high level” math students are matched with “low level” math students or varying combinations of “math levels” and “math grades” on teams. Does team structure based on math level or math grade in the FtC classroom positively impact the concept learning of low math level students?
BIBLIOGRAPHY


Table 6: Least Squares Means Matrix and Confidence Intervals for 3-Way Interaction Model 1

<table>
<thead>
<tr>
<th>Mean[j] Buffer</th>
<th>Mean[j] Buffer</th>
<th>Mean[j] Buffer</th>
<th>Mean[j] Buffer</th>
<th>Mean[j] Buffer</th>
<th>Mean[j] Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>Mean[j] Buffer</td>
<td>Mean[j] Buffer</td>
<td>Mean[j] Buffer</td>
<td>Mean[j] Buffer</td>
<td>Mean[j] Buffer</td>
<td>Mean[j] Buffer</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
<td>1, 2, 3</td>
</tr>
</tbody>
</table>

354
Table 7: Least Square Means Matrix and Confidence Limits for 3-Way Interaction Model 2

<table>
<thead>
<tr>
<th>Mean[i]-Mean[j]</th>
<th>Std Err Df</th>
<th>Lower CL Df</th>
<th>Upper CL Df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1,2,3</td>
<td>-0.0415</td>
<td>-0.0415</td>
<td>-0.0415</td>
</tr>
<tr>
<td>1,2,4</td>
<td>0.00205</td>
<td>0.00205</td>
<td>0.00205</td>
</tr>
<tr>
<td>1,3,2</td>
<td>0.09489</td>
<td>0.09489</td>
<td>0.09489</td>
</tr>
<tr>
<td>1,3,3</td>
<td>0.56252</td>
<td>0.56252</td>
<td>0.56252</td>
</tr>
<tr>
<td>1,3,4</td>
<td>3.12932</td>
<td>3.12932</td>
<td>3.12932</td>
</tr>
<tr>
<td>2,2,2</td>
<td>1.29072</td>
<td>1.29072</td>
<td>1.29072</td>
</tr>
<tr>
<td>2,2,3</td>
<td>3.19872</td>
<td>3.19872</td>
<td>3.19872</td>
</tr>
<tr>
<td>2,2,4</td>
<td>2.36866</td>
<td>2.36866</td>
<td>2.36866</td>
</tr>
<tr>
<td>2,3,2</td>
<td>0.00973</td>
<td>0.00973</td>
<td>0.00973</td>
</tr>
<tr>
<td>2,3,3</td>
<td>0.00973</td>
<td>0.00973</td>
<td>0.00973</td>
</tr>
<tr>
<td>2,3,4</td>
<td>0.00973</td>
<td>0.00973</td>
<td>0.00973</td>
</tr>
</tbody>
</table>

LSMean[i]