

UNIVERSITY OF OKLAHOMA  
GRADUATE COLLEGE

APPLICATION OF COGNITIVE PRINCIPLES WITHIN AN ONLINE  
STATISTICAL LEARNING ENVIRONMENT

A DISSERTATION  
SUBMITTED TO THE GRADUATE FACULTY  
in partial fulfillment of the requirements for the  
Degree of  
DOCTOR OF PHILOSOPHY

By  
WILLIAM B. HUFFMAN  
Norman, Oklahoma  
2016

APPLICATION OF COGNITIVE PRINCIPLES WITHIN AN ONLINE  
STATISTICAL LEARNING ENVIRONMENT

A DISSERTATION APPROVED FOR THE  
DEPARTMENT OF PSYCHOLOGY

BY

---

Dr. Sowon Hahn, Chair

---

Dr. Robert Terry

---

Dr. Scott Gronlund

---

Dr. Lauren Ethridge

---

Dr. Barbara Greene

© Copyright by WILLIAM B. HUFFMAN 2016  
All Rights Reserved.

## **Acknowledgements**

First, I would like to thank my mother and grandfather for their unwavering support and continuous teachings. Second, I would like to thank my master's thesis committee from the University of Central Oklahoma, in particular Dr. Thomas Hancock for his superb guidance and instruction.

Next, I thank my dissertation committee here at the University of Oklahoma, their input and feedback is indeed what made this project a success. Dissertation committee members included: Dr. Sowon Hahn, Dr. Robert Terry, Dr. Scott Gronlund, Dr. Lauren Ethridge, and Dr. Barbara Greene, as well as Dr. Taehun Lee.

Lastly, I want to thank my committee chair Sowon Hahn again for her massive amount of support and continued mentorship. It truly is amazing what can happen when a few talented individuals set out to do something awesome! You all have my thanks!

## Table of Contents

Acknowledgements .....	iv
List of Tables .....	vi
List of Figures.....	vii
Abstract.....	viii
Chapter 1: Introduction.....	1
Chapter 2: Optimal Learning Phenomena.....	3
Chapter 3: Experiments.....	9
Experiment 1.....	11
Experiment 2.....	29
Experiment 3.....	46
Chapter 4: Item Response Theory.....	62
Chapter 5: Application of Item Response Theory.....	70
Experiment 1.....	71
Experiments 2 & 3.....	73
Chapter 6: General Discussion.....	77
References.....	83
Appendix A: Statistics Test.....	86
Appendix B: Transfer Test.....	90

## List of Tables

Table 1. Overview of Hypotheses.....	11
Table 2. Demographics for Experiment 1.....	12
Table 3. Experiment 1 Inter-rater Reliability.....	22
Table 4. Participant's Preference and Best Predicted for Learning.....	27
Table 5. Demographics for Experiment 2.....	30
Table 6. Experiment 2 Inter-rater Reliability.....	33
Table 7. Experiment 2 Transfer Inter-rater Reliability.....	34
Table 8. Experiment 2 Participant's Preference and Best Predicted for Learning.....	43
Table 9. Experiments 1 & 2 Learning Condition Performance by Final Test Format..	44
Table 10. Demographics for Experiment 3.....	46
Table 11. Experiment 3 Inter-rater Reliability.....	48
Table 12. Experiment 3 Transfer Inter-rater Reliability.....	49
Table 13. Experiment 3 Participant's Preference and Best Predicted for Learning.....	59
Table 14. Two-Parameter-Logistical Model for Statistical Questions in Experiment 1.....	71
Table 15. Two-Parameter-Logistical Model for Transfer Questions in Experiments 2 & 3.....	73
Table 16. Outcome of Project Hypotheses.....	77

## List of Figures

Figure 1. Multiple Evaluation Example.....	15
Figure 2. Procedural Flow Chart.....	17
Figure 3. Multivariate Results of Experiment 1.....	21
Figure 4. Correctly Answered Open Ended Questions by Learning Condition.....	35
Figure 5. Correctly Answered Multiple Choice Questions by Learning Condition.....	37
Figure 6. Correctly Answered Multiple Evaluation Questions by Learning Condition.....	39
Figure 7. Correctly Answered Open Ended Questions by Feedback and Exam Type..	51
Figure 8. Correctly Answered Multiple Choice Questions by Feedback and Segmentation Type.....	52
Figure 9. Correctly Answered Multiple Choice Questions by Learning Condition and Exam Type.....	53
Figure 10. Correctly Answered Multiple Evaluation Transfer Questions by Learning Condition.....	55
Figure 11. Correctly Answered Multiple Evaluation Questions by Learning Condition and Exam Type.....	56
Figure 12. One Latent Factor with 20 Predictors Model.....	62
Figure 13. Two-Parameter-Logistical Item Response Theory Model with Varying Item Difficulties.....	64
Figure 14. Two-Parameter-Logistical Item Response Theory Model with Varying Item Discriminations.....	66
Figure 15. Hypothetical Item Difficulties for a "Good" Test.....	68
Figure 16. Hypothetical Item Discriminations for a "Good" Test.....	69
Figure 17. Item Characteristic Curves for the Original Statistical Test.....	72
Figure 18. Item Characteristic Curves for the Transfer Test.....	75

## **Abstract**

Three experiments were conducted in order to further investigate optimal learning procedures within an online statistical learning environment. Experiment 1 exposed learners to retrieval practice learning conditions with or without segmentation. Retrieval practice formats included; multiple choice, open ended, multiple evaluation, or instruction only type manipulations. Experiment 2 explored the impact of added immediate feedback in conjunction with retrieval practice and segmentation. Experiment 3 further investigated how the benefits of optimal learning procedures transfer to novel situations / examinations. Within all experiments a series of metacognitive questions were administered to learners in order to measure their metamemory over the statistical knowledge that was taught. In alignment with our hypotheses and previous research it was found that retrieval practice (experiment 1) and retrieval practice with immediate feedback (experiment 2) tended to boost memory retention. However, the data trend for all experiments tended to suggest that segmentation has little or no impact on statistical learning, such a finding was support against our hypotheses as well as the findings within previous studies. Though the results are somewhat mixed, the benefits associated with retrieval practice and retrieval practice with feedback did not seem to transfer to novel instances. Individuals that learned within an open ended or multiple evaluation type format tended to have greater insight into their own metacognitive knowledge.

*Keywords:* optimal learning, retrieval practice, feedback, segmentation effect, multiple evaluations, metamemory



## Chapter 1: Introduction

As technology continues to advance, so does the instructors ability to utilize technology within classroom environments effectively. Fully electronic classroom environments are now possible and are being used at universities around the world. Online classrooms may provide us with powerful tools to teach with and at the same time allow students the ability to attend courses they otherwise may not have been able to (due to time or location). However, whether or not the online classroom is effective as a traditional classroom seems dependent on the subject material being taught (Kemp & Grieve, 2014; Maki & Maki, 2002). For example, at the University of Oklahoma individuals that are instructed in introduction to statistics often struggle when the course is online compared to when it is administered within a classroom. The difference is about a full letter grade favoring the traditional method of instruction. The low online statistics scores obtained by students in introduction to statistics was one of the motivations for this project. An immersive online multimedia instructional environment with cognitive learning and memory enhancement theory embedded was developed at an attempt to better instruct future statistical learners.

Though automated instruction has been around since the late 1950's with Skinner's teaching machines, the technology available today as well as the advancements in cognitive learning theories suggests that additional research must be conducted in order to find optimal instructional methods within online learning environments. Multimedia instruction is not a new idea however; as cognitive psychologists and instructors utilizing technology we must make sure that we are constructing effective learning apparatus' that include a foundation of cognitive

principles. A goal of this project is to develop an effective and efficient system of multimedia learning while keeping sound cognitive principles in mind. An overview of some of the common cognitive learning procedures will be reviewed in chapter 2.

This project also has three goals. First, this project will investigate known effective learning procedures in modern multimedia learning environments utilizing complex statistical stimuli. Second, this project will also look into the effectiveness of multiple evaluations to be used as a tool for learning (not just a tool for measurement) within a multimedia learning environment, an endeavor that has yet to be investigated. And, third this project will further investigate unknown optimal learning procedures, such as the interactive effects of retrieval practice and the segmentation effect. In order to find new and better ways to instruct future students, such optimal learning procedural experiments must be conducted.

## **Chapter 2: Optimal Learning Phenomena**

### **Retrieval Practice**

Retrieval practice is the finding in which initial learning paired with an immediate test often results in greater long term memory than additional study (Roediger & Karpicke, 2006). Retrieval practice has been found to be beneficial for scientific learning courses that were presented in a multimedia type fashion (Johnson & Mayer, 2009). The phenomenon has even seen effectiveness in challenging areas of learning such as foreign language learning where the learners have little or no prior knowledge of the topic in which they learn (Kang, 2010; Huffman & Hahn, 2015). In fact, Rowland (2014), conducted a large metaanalysis in which the retrieval practice phenomenon was investigated. The researcher found that the majority of experiments yielded medium to large effect sizes within a wide array of areas of memory and learning.

Many memory researchers suggest that retrieval practice works by strengthening pathways to target to-be-remembered information (Halamish & Bjork, 2011; Pyc & Rawson, 2009), thereby lessening the likelihood of forgetting (Kornell, Bjork, & Garcia, 2011). Many researchers have also investigated the impacted of repeated testing (Huffman & Hahn, 2015; Karpicke & Roediger, 2007; Larsen, Butler, & Roediger, 2009). Though multiple instances of retrieval practice seem better than one, the first correct retrieval seems to be the most beneficial for learning (Pyc & Rawson, 2009).

Though the retrieval practice phenomenon is robust enough to be used in isolation (Rowland, 2014), this project will build on the idea of optimal learning, the idea of combining multiple learning methods in order to find new effective instructional

methods (Maddox & Balota, 2015). For example Fritz, Morris, Acton, Voelkel, Etkind (2007) investigated the keyword method in conjunction with retrieval practice; and though there was not an additive or multiplicative impact on learning to be found when keyword and retrieval practice methodologies were combined into one procedure, it was still necessary to be conducted to see if it would have been beneficial for learning. One optimal learning method that has seen some promise is the spaced testing effect (Carpenter & DeLosh, 2005; Maddox & Balota, 2015). Essentially, the spaced testing effect is a combination of both retrieval practice and the spacing effect phenomena in which learners tend to receive an additive boost from both types of learning procedures. This project will seek to further investigate novel optimal learning procedures.

Retrieval practice will be utilized in this experiment to see if it can boost statistical learning within an electronic environment. We planned to implement this procedure directly by teaching students new information and then immediately testing them over that respective information. Should retrieval practice be beneficial, added manipulations will be placed into the experiment in order to make the impact of retrieval practice even more effective on participant's learning. One previously known way to boost the impact of retrieval practice even further is through the use of Feedback (Butler & Roediger, 2008).

### **Feedback**

Though retrieval practice has been found to be effective on its own (Pyc & Rawson, 2009), research shows that feedback can increase the impact of learning of to-be-learned information (Butler & Roediger, 2008). Feedback can be implemented effectively either immediately (Hayes, Kornell, & Bjork, 2010) or at a delay (Smith &

Kimball, 2010). Which type of feedback is better typically depends on the difficulty of the test. If you (as a teacher) think learners are going to correctly answer the question the first time around, feedback is typically better at a delay (Smith & Kimball, 2010). However, if you think learners are going to get the question incorrect, feedback is typically better allocated if done immediately (Hayes et al., 2010). Because retrieval practice is a powerful memory mechanism, once the answer to a question is recalled learners tend to remember that respective recall regardless if it is correct or incorrect (McDermott, 2006). When learners miss or fail the first recall immediate feedback tends to be more beneficial for the learner because the respective memory trace is still active in memory and can be corrected if need be (Hayes, Kornell, & Bjork, 2013).

Because statistics is often found to be challenging for novice learners, immediate feedback will be utilized within this experiments design. Feedback can be given in multiple fashions; however corrective feedback will be utilized within our procedure due to some conditions having open ended type questions. Corrective feedback functions by providing feedback to learners regardless of whether they answered the question correctly or not. The corrective feedback is expected to provide the learners with two things (Butler & Roediger, 2008). First and foremost it will provide learners with the correct answer, and secondly should the learner get the question correct it is expected to increase the learner's confidence of that correct answer. It is expected that tests with immediate feedback conditions will behove learning within this study more so than test only conditions (Butler, Karpicke, & Roediger, 2007).

## **Segmentation**

Segmentation is an instructional tool that can be utilized within a multimedia environment by allowing learners to encode new information at their own pace (Mayer, 2008). Researchers found that when learning new information, participants often retained information better when it was presented in learner paced chunks (segmentation) rather than having the new information presented to them continuously (Mayer, Mahias, & Wetzell, 2002). Because our encoding processes are limited when presented with new information (Baddeley, 1992; Miller, 1956), the segmentation phenomenon is expected to help reduce cognitive load in an effort to allow students to learn complex material effectively at their own pace. Mayer's work tends to suggest that segmentation procedure works best within user driven environments.

Segmentation will be utilized in this experiment in conjunction with retrieval practice and retrieval practice with feedback. This interaction is a novelty to this project that has yet to be investigated. Segmentation conditions will be implemented by allowing learners to stop after 2 minutes of instruction and then answer questions. When the learner has completed the questions and feel that they are ready to proceed they may do so. This procedure will continue throughout the study until all instruction and questions are completed. It should be noted that this segmentation procedure will be manufactured by the instructor not the students. It is expected that applying both retrieval practice in conjunction with segmentation may result in added memory benefits due to the utilization of two different learning theories simultaneously.

## **Metamemory**

Metamemory is the knowledge that we have access to about our own memories. Within a learning context this could be confidence of a correct answer on an exam or a general overall knowledge of how one performed on a test administered. Many times these types of questions are referred to as judgments of learning (or JOLs). It is a common finding in cognitive learning experiments that students are often overconfident about their respective memories when encoding new material (Kornell & Bjork, 2009). Dirkwager (1996) proposes the idea of personal probabilities (or multiple evaluations) in which, when calibrated correctly, should result in less guessing thereby improving learners respective judgments of learning. Though JOLs are often studied by cognitive scientists they are rarely studied in conjunction with optimal learning procedures that will be utilized within the current project.

Multiple evaluations are typically thought of as a probabilistic type of measurement (Dirkwager, 2003). Multiple evaluations are similar to multiple choice questions but instead of asking learners to choose an absolute answer, learners must rate each answer by providing a confidence rating (see Figure 1 for an example). Providing learners with a self-monitoring procedure like multiple evaluations or similar types of JOLs has been found to improve student's metamemory insight (Dunlosky, Hertzog, Kennedy, & Thiede, 2005). Because multiple evaluations are essentially the same as a multiple choice test in terms of how they are administered, multiple evaluations will be able to be implemented within this project directly as a type of question that can be manipulated between conditions. Evidence suggests that individuals that are administered multiple evaluations while learning will have greater insight into their own

knowledge and thus be able to better predict their own performance (Dunlosky et al., 2005). Though it is suggested that multiple evaluations provide better precision in testing ones knowledge than multiple choice tests (Dirkzwager, 2003), this project will also look into effectiveness of multiple evaluations to be used as a learning tool within a multimedia learning environments.



## Chapter 3: Experiments

The main goal of all experiments within this project is to implement effective instruction and learning of statistical material within an online multimedia environment. In Experiment 1 participants will be randomly assigned to one of seven learning conditions; open ended questions that are segmented or unsegmented , multiple choice questions that are segmented or unsegmented, multiple evaluation questions that are segmented or unsegmented, or an instruction only control. Experiment 2 will implement the same 6 experimental conditions administered with immediate corrective feedback as well as a control group that receives no questions. Experiment two will also inquire as to how well knowledge transfers to a similar, but different, quiz. Experiment 3 is essentially a replication of experiment 1 except with added transfer questions to assess knowledge transfer.

Dependent variables included are the results of a 20 question quiz and a 20 question transfer quiz that will be administered after a 24-hour delay. A series of metacognitive questions will also be asked during part 2 of each experiment to assess learner's metacognitive knowledge. It is hypothesized that retrieval practice, retrieval practice with feedback, and segmentation will all be beneficial for learning. It is hypothesized that retrieval practice along with segmentation together will create an additive interaction so that the learner will benefit from both types of learning methods. Because multiple evaluations are engaging learners in a judgment of learning type task, it is expected that learners that are assigned to this condition will have a greater metacognitive insight into their own knowledge than individuals that were taught in a different fashion. Lastly it is expected that retrieval practice, retrieval practice with

feedback, and segmentation conditions will all outperform the control condition on both the original statistics and the knowledge transfer quiz.

Multiple evaluations have a unique scoring system that will be mimicked in this project (for the original use of how the multiple evaluation scoring system works see Dirkzwager, 2003). Because the original penalty function used by Dirkzwager could not be directly programmed into Qualtrics, the score weights were hard coded. Correct answers were weighted with a 1 and incorrect answers were weighted with a  $(-1/3)$ , thus allowing for the scoring system in this project to act similar to the function used by Dirkzwager. For example individuals that are highly confident on the correct answer received full credit (100 points). Individuals that are highly confident on the incorrect answer were penalized (-33 points). Individuals that are unsure of the correct answer and that put 25% on all possible solutions received no credit or penalty (0 points) Individuals that are convinced that questions 50% A or 50% B are correct, and one of them is correct received partial credit (about 33.33 points). Weighting the points in this fashion did not mimic Dirkzwager's penalty function perfectly but it was fairly close. A list of all main hypotheses can be seen in table 1.

## **Table 1. Overview of Hypotheses**

---

### **Experiment 1**

---

- Retrieval practice conditions will outperform the instruction only condition.
  - Open ended and multiple evaluation formats will behoove learning more than a multiple choice format.
  - Conditions that are segmented will outperform conditions that are not segmented.
  - An additive learning effect between the retrieval practice and segmentation procedures are expected.
  - Multiple evaluation type formats will have greater insight into their own knowledge.
- 

### **Experiment 2**

---

- Retrieval practice with feedback conditions will outperform retrieval practice and instruction only conditions.
  - Conditions that are segmented with feedback will outperform conditions that are not segmented.
  - An additive learning effect between the retrieval practice and segmentation procedures are expected.
  - Retrieval practice with feedback conditions will outperform retrieval practice and instruction only conditions in terms of transfer.
  - The added manipulation of corrective feedback will boost all learners' insight into their own knowledge.
- 

### **Experiment 3**

---

- Retrieval practice conditions will outperform the instruction only condition in terms of transfer knowledge.
  - Conditions that are segmented will outperform conditions that are not segmented in terms of transfer.
  - Multiple evaluation type formats will have greater insight into their own transfer knowledge.
- 

## **Experiment 1**

### **Method**

### **Participants**

Two-hundred and Twenty students from the University of Oklahoma volunteered for Experiment 1. Fifty-eight participants were excluded from the data analyses for failing to return to part two of the experiment. Of the remaining 158 participants, 118 were female with ages ( $M = 19.46$ ,  $SD = 3.14$ ) ranging from 18 to 41. Class level reported indicated 69.8% freshmen, 22.8% sophomores, 6.2% juniors, and 1.2% seniors. Participants reported an average GPA of ( $M = 3.36$ ,  $SD = 0.50$ ) and an ACT score of ( $M = 26.49$ ,  $SD = 4.27$ ). Additional descriptive statistics are listed in Table 2.

**Table 2. Demographics for Experiment 1**

	<b>Count</b>	<b>Percentage</b>
<b>Statistics Knowledge</b>		
I have not taken a class nor do I have any prior experience with formal statistics	96	59.3%
I have not taken a class but I do have some experience with formal statistics	10	6.2%
I have taken a class at the high school level	30	18.5%
I have taken a class at the college level	26	16%
I have taken a class at the graduate level	0	0%
<b>Research Methods Knowledge</b>		
I have not taken a class nor do I have any prior experience with research methods	110	67.9%
I have not taken a class but I do have some experience with research methods	32	19.8%
I have taken a class at the high school level	10	6.2%
I have taken a class at the college level	10	6.2%
I have taken a class at the graduate level	0	0%
<b>Math Knowledge</b>		
High school mathematics	19	11.7%
Some college mathematics	35	21.6%
College algebra	38	23.5%
College calculus	65	40.1%

---

**Materials**

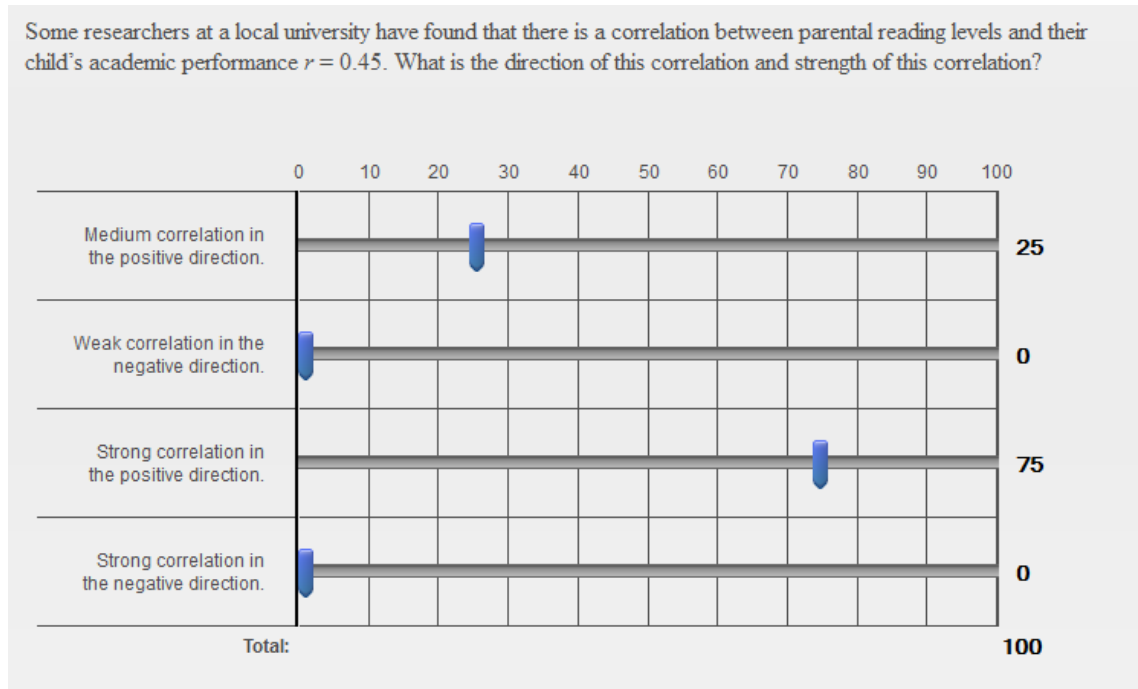
Twenty potential test questions for PSY 2003 / PSY 3114 were developed. The topics of these questions were as follows: correlations, measurement, research design, and sampling questions. Five questions were developed within each subsection topic. These 20 questions were developed into 3 different question formats (open ended, multiple choice, and multiple evaluations). A 20 minute instructional video was also developed in order provide instruction within an online environment in a standardized fashion. The instructional videos were made by the primary investigator and can be found at this URL: <https://www.youtube.com/watch?v=NYrrKydPsN4> . Though the video was the same across all conditions within this experiment, the video listed above was cropped into multiple formats for the respective manipulation and to partially counter balance the conditions in which participants were assigned to groups. Though order effects were considered to be unlikely, multiple orders of the videos 4 main topics were constructed. The program itself was made using Qualtrics, an online data collection tool that allows for psychometric tools to be implemented. Participants viewed the program on 17" square monitors and listened to the audio through Sony circumaural headphones.

## **Procedure and Design**

All participants were recruited from University of Oklahoma's online participant's pool (SONA). Individuals were told they would be watching an educational video and answering questions during or at the end of the experiment on day 1. Each participant was randomly assigned to one of seven conditions. Before the experiment began, learners were asked a series of demographic information. This project implemented a 3 (question type) x 2 (segmentation type) between design with an added control condition. The question type factor consisted of how the learner was tested during the learning phase of the experiment (open ended, multiple choice, or multiple evaluation). The control condition received no questions during the learning phase of the experiment. The segmentation factor consisted of whether or not that instruction was segmented or unsegmented.

In the open ended condition participants were asked questions in a standard open ended format and in the multiple choice condition participants were asked questions during the learning phase in a typical 4 answer possible multiple choice format. For the multiple evaluation condition participants were given a question and 4 possible answers with an added judgment of learning task of confidence. Learners were asked to rate how confident the respective answer was correct for all answers. If they were highly confident that B was the correct answer they were instructed to put B as 100% and all other choices as a 0%. If the participants were sure it was a or c they were instructed to put 50% on a and 50% on c. If they were unsure of the correct answer they were instructed to put 25% on all choices. The participants were told that highly confident

incorrect answers would result in a deduction of points. Fifty percent on the correct response and 50% on an incorrect response would result in partial credit and 100% on the correct answer would result in full credit for that question (see Figure 1 for an example).



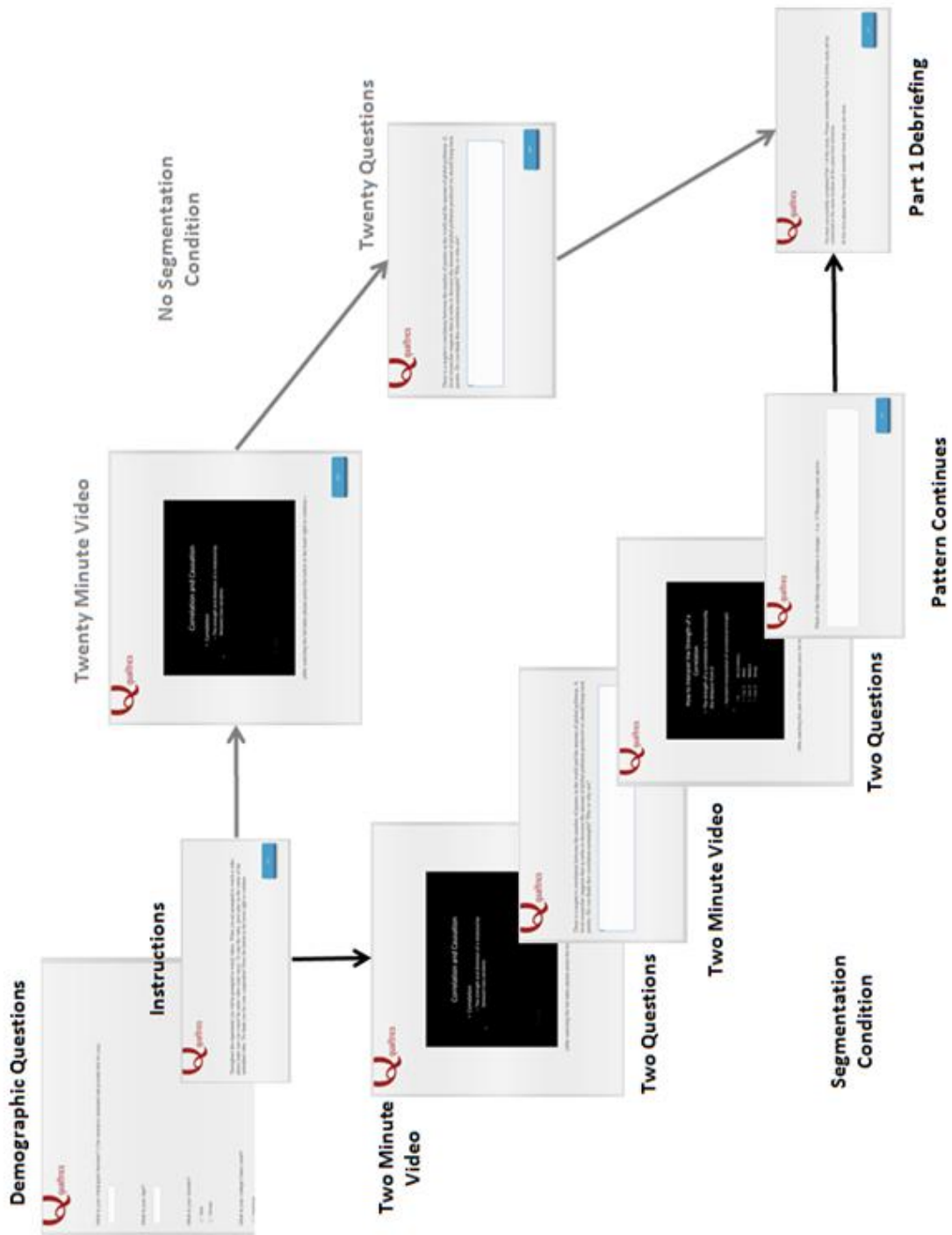
**Figure 2. Multiple Evaluation Example**

For the segmentation manipulation individuals were presented with 2 minutes of instruction followed by 2 questions (on average) during the learning phase of the experiment. This process continued until all 20 minutes of the instructional video was seen and all 20 questions were administered. In the unsegmented conditions participants watched 20 minutes of instruction and then answered the 20 question statistical battery.

Demographic information included age, gender, class level (freshmen, sophomore, etc.), current GPA, ACT score submitted to the University of Oklahoma, prior knowledge of statistics, research knowledge, and math. Prior knowledge questions (statistics and research methods) were asked in a multiple choice format with the following choices: "I have not taken a class nor do I have any prior experience with formal statistics," "I have not taken a class but I do have some experience with formal statistics," "I have taken a class at the high school level," "I have taken a class at the college level," or "I have taken a class at the graduate level." All participants were asked to count concurrent enrollment. The research methods prior knowledge question used the above format but substituted the word statistics for the phrase research methods. The math prior experience question listed the following possibilities: "high school mathematics," "some college mathematics," "college algebra," "college calculus," or "college linear algebra or higher." All of the demographic information collected were used as potential covariates within the data analyses conducted in this study.

After the demographic information was submitted participants continued within their respective condition of their experiment until they were completed watching all 20 minutes of instruction and after answering all 20 questions. No feedback was administered during experiment 1. After Part 1 of the experiment was complete participants were given course credit, reminded about part 2 of the experiment, and thanked for their time. Part 1 lasted approximately 45 minutes (see Figure 2 for a procedure flow chart).





**Figure 2. Procedural Flow Chart**

Part 2 of the experiment was conducted after a 24 hour delay. The dependent variables of interest within this experiment are the 20 questions in 3 different test formats. In experiment 1 participants answered 20 questions within all 3 formats (60

total) in the following order: open ended, multiple evaluation, and then multiple choice. This order of format was kept consistent across all learners in order to reduce carry over effects. Again, no feedback was given during experiment 1 After each 20 question test learners were asked a series of metacognitive questions. Metacognitive questions consisted of a prediction of their own performance (how well did you think you did on the test that was just administered) and a difficulty judgment about the format of the test (how hard was this test in this format on a Likert scale, with a 7 indicating very difficult and a 1 indicating very easy). After all 60 questions were administered 2 additional metacognitive questions were asked. These questions asked the learners to tell us which format they liked the best and which format administered during the learning phase would result in the best learning. Part 2 of the experiment took about 30 minutes to complete. When learners were completed with the experiment they were debriefed, given course credit, and thanked for their time.

It was hypothesized that questions in any question type format would outperform the no question control group. In addition it was also predicted that the open ended questions and multiple evaluation conditions would outperform the multiple choice questions due to those respective formats being more cognitively effortful. In accordance with Mayer's work it is suggested that segmentation of learning information should behoove learning more so than presenting the information in an unsegmented fashion (Mayer & Moreno, 2003; Mayer, 2008). Due to the added metacognitive benefits of the multiple evaluation questions it is expected that learners that are asked questions in this fashion will have greater insight into their own knowledge and therefore will be able to better predict their own performance.

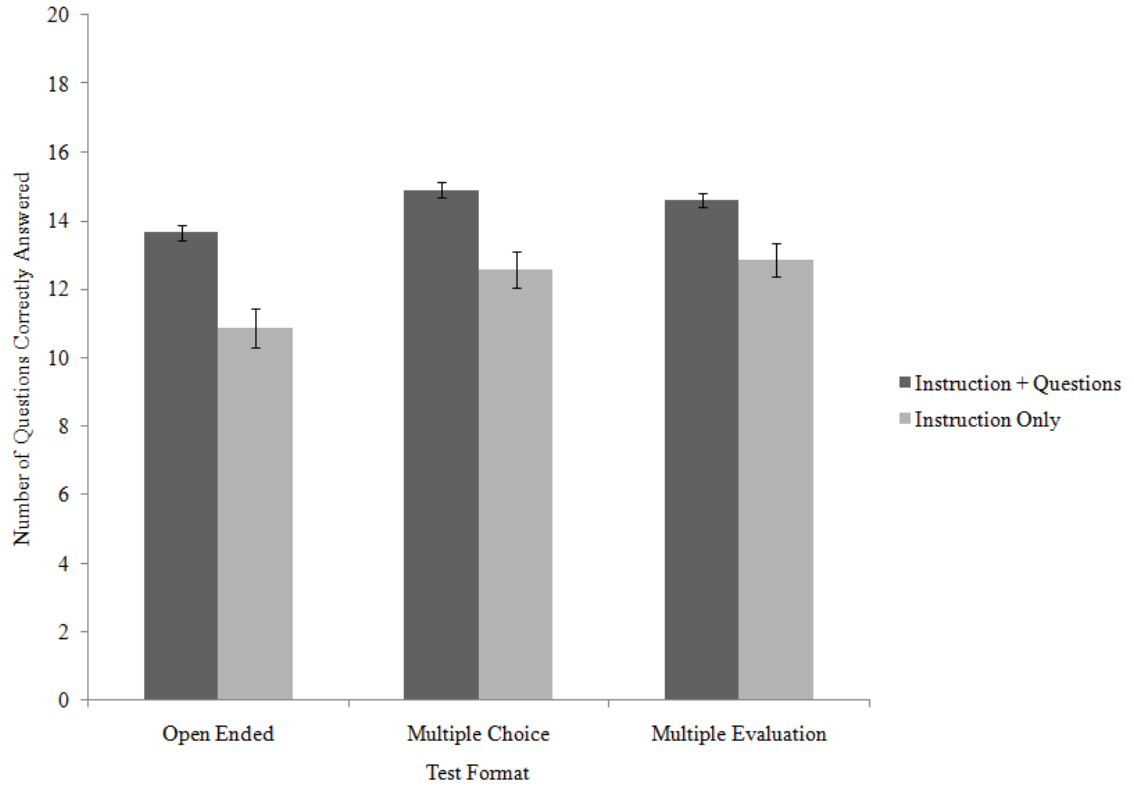
## Results (Experiment 1)

### DV: Multivariate

An initial MANCOVA was conducted and yielded that the only significant covariate was ACT score; all other covariates were dropped from the analysis. A 3 (question type) X 2 (segmentation type) MANCOVA was run with ACT score as a covariate. Dependent measures within this analysis included open ended answers, multiple choice answers, and multiple evaluation answers. At the multivariate level there was no interaction found between question and segmentation type factors (Wilks'  $\lambda = 0.98$ ,  $F(6, 212) = 0.37$ ), nor a main effect for question type (Wilks'  $\lambda = 0.94$ ,  $F(6, 212) = 1.14$ ), or segmentation type (Wilks'  $\lambda = 0.996$ ,  $F(6, 89) = 0.14$ ). ACT score (Wilks'  $\lambda = .64$ ,  $F(3, 106) = 19.76$ ,  $p < 0.001$ ,  $d = 1.50$ ) was a significant covariate within the analysis. Individuals with higher ACT scores tended to do better on the final statistical examination regardless of which condition they were randomly assigned to.

At the univariate level no interaction was found between question type and segmentation type for open ended, multiple choice, or multiple evaluation answers. No main effect of question type or segmentation type was found between any answer types. At the univariate level ACT score (open ended  $F(1, 108) = 58.33$ ,  $p < 0.001$ ,  $d = 1.47$ , multiple choice  $F(1, 108) = 48.43$ ,  $p < 0.001$ ,  $d = 1.34$ , and multiple evaluation  $F(1, 108) = 42.78$ ,  $p < 0.001$ ,  $d = 1.25$ ) continued to be a significant covariate across all question types.

An additional MANCOVA was conducted including the control group with ACT score as a covariate. This specific analysis was looking at all tests types administered during the learning phase vs no testing during the learning phase of the experiment. A main effect of question type was found (Wilks'  $\lambda = 0.86$ ,  $F(3, 129) = 7.23$ ,  $p < 0.001$ ,  $d = 0.81$ ) such that individuals that were administered a test during phase 1 of the experiment did better than those individuals not administered a test during the learning phase. ACT (Wilks'  $\lambda = 0.69$ ,  $F(3, 129) = 18.97$ ,  $p < 0.001$ ,  $d = 1.34$ ) was a significant covariate. Again individuals with higher ACT scores tended to better on the final statistics examination regardless of which condition they were randomly assigned to. At the univariate level the same data pattern was seen for question type across all dependent measures (open ended  $F(1, 131) = 20.53$ ,  $p < 0.001$ ,  $d = 0.81$ , multiple choice  $F(1, 131) = 17.90$ ,  $p < 0.001$ ,  $d = 0.74$ , and multiple evaluation  $F(1, 131) = 11.70$ ,  $p < 0.001$ ,  $d = 0.59$ ). Individuals tested during the learning phase (open ended:  $M = 13.65$ ,  $SD = 1.59$ ; multiple choice:  $M = 14.90$ ,  $SD = 1.44$ ; multiple evaluation:  $M = 1460.03$ ,  $SD = 134.19$ ) did better than individuals that were not tested during the learning phase (open ended:  $M = 10.85$ ,  $SD = 2.73$ ; multiple choice:  $M = 12.57$ ,  $SD = 2.45$ ; multiple evaluation:  $M = 1276.28$ ,  $SD = 238.74$ ) across all dependent measures (see Figure 3, error bars in all figures within this document represent standard error).



**Figure 3. Multivariate Results of Experiment 1**

**DV: Open Ended**

The open ended questions were graded by 3 independently trained research assistants. As with all dependent variables in this study, the open ended format of the test consisted of 20 questions ( $\alpha = 0.95$ ). A breakdown of each individual item's interrater reliability can be seen in Table 3.

**Table 3. Experiment 1 Inter-rater Reliability**

<b>Question Number</b>	<b>Cronbach's <math>\alpha</math></b>
1	0.96
2	0.99
3	0.99
4	0.97
5	0.98
6	0.98
7	0.70
8	0.78
9	0.96
10	0.95
11	0.98
12	0.88
13	0.91
14	0.98
15	0.98
16	0.89
17	0.98
18	0.77
19	0.92
20	0.92
Total	0.95

An initial ANCOVA suggested that the only significant covariate was ACT score. A 3 (question type) X 2 (segmentation type) ANCOVA was conducted with ACT score as a covariate. No interaction between question type and segmentation type was found  $F(2, 115) = 0.23$ . No main effect of question type was found  $F(2, 115) = 1.64$ , nor segmentation type  $F(2, 115) = 0.26$ . ACT score continued to be a significant covariate ( $F(1, 115) = 58.33, p < 0.001, d = 1.47$ ) for the model. The significant covariate implies that individuals with a high ACT score tended to well on the final statistics examination regardless of which condition they were randomly assigned to.

An additional ANCOVA was conducted with the control condition; ACT score remained as a covariate. A main effect was found for the question type factor ( $F(1, 131) = 20.53, p < 0.001, d = 0.81$ ) such that questions asked during the instructional

phase ( $M = 13.65$ ,  $SD = 2.71$ ) benefited learners more so than if no questions ( $M = 10.85$ ,  $SD = 2.73$ ) were given (see Figure 3). ACT score remained a significant covariate ( $F(1, 131) = 56.50$ ,  $p < 0.001$ ,  $d = 1.31$ ) suggesting that individuals with higher ACT scores tended to do better on the final statistical examination within an open ended format.

### **DV: Multiple Choice**

An initial ANCOVA suggested that the only significant covariates were GPA and ACT score for the multiple choice dependent variable so all other were dropped from this part of the analysis. A 3 (question type) X 2 (segmentation type) ANCOVA was conducted. No interaction was found between question type and segmentation type factors  $F(2, 97) = 0.316$ . No main effects were found between question types  $F(2, 97) = 1.25$  or segmentation type  $F(1, 97) = 1.55$  factors. Significant covariates included GPA  $F(1, 97) = 5.30$ ,  $p = 0.02$ ,  $d = 0.46$  and ACT scores  $F(1, 97) = 28.10$ ,  $p < 0.001$ ,  $d = 1.09$ . According to the data, higher GPAs and higher ACT scores seem to be having a positive impact on correct multiple choice scores regardless of the condition in which they were randomly assigned.

An additional one-way ANCOVA was conducted including the control condition. It was found that asking questions ( $M = 14.78$ ,  $SD = 2.59$ ) during the learning phase of the experiment was more beneficial than not asking questions ( $M = 12.83$ ,  $SD = 2.64$ ) during the learning phase of the experiment in terms of long term retention in a multiple choice format  $F(1, 118) = 10.96$ ,  $p = 0.001$ ,  $d = 0.63$  (see Figure 3).

Significant covariates included GPA  $F(1, 118) = 5.96, p = 0.02, d = 0.46$  and ACT scores  $F(1, 118) = 27.52, p < 0.001, d = 0.97$ . For multiple choice questions it seemed that high GPA and ACT score tended to boost final statistical performance regardless of what condition participants were assigned.

### **DV: Multiple Evaluations**

An initial ANCOVA suggested that the only significant covariates were GPA and ACT score for the multiple evaluation format so all other were dropped from this part of the analysis. A 3 (question type) X 2 (segmentation type) ANCOVA was conducted for multiple evaluation questions correctly answered (raw evaluation scores that were correct were used for this part of the analysis). No penalties were included for the multiple evaluation scores in the present analysis. The ANCOVA yielded no significant interaction between question type and segmentation type factors  $F(2, 97) = 0.46$ . No main effects were found for the question type  $F(2, 97) = 0.147$  or segmentation type  $F(1, 97) = 0.38$  conditions in terms of correctly answered multiple evaluation questions. Significant covariates included GPA  $F(1, 97) = 4.18, p = 0.045, d = 0.41$  and ACT scores  $F(1, 97) = 21.42, p < 0.001, d = 0.94$ . Again, individuals with higher GPAs and ACT scores reported tended to score higher on the multiple evaluation test than individuals with lower GPAs and ACT score regardless of the learning condition that were in.

An additional one-way ANCOVA was conducted including the control condition. It was found that individuals that were administered questions ( $M = 1455.18$ ,



$SD = 249.36$ ) during the learning phase significantly outperformed individuals that were not administered questions ( $M = 1284.81$ ,  $SD = 254.76$ ) in terms of multiple evaluation questions correctly answered  $F(1, 118) = 10.96$ ,  $p = 0.001$ ,  $d = 0.63$  (see Figure 3). Significant covariates included GPA  $F(1, 118) = 4.48$ ,  $p = 0.04$ ,  $d = 0.41$  and ACT scores  $F(1, 118) = 19.40$ ,  $p < 0.001$ ,  $d = 0.81$ .

The same two ANCOVA's were computed for adjusted Dirkzwager scores. An initial ANCOVA suggested that ACT score and GPA were the only significant covariates so all other were dropped from the data analysis. No interaction was found between question type and segmentation type factors  $F(2, 97) = 0.46$ . Nor was there a main effect of question type  $F(2, 97) = 0.15$  or segmentation type  $F(1, 97) = 0.38$ . ACT score  $F(1, 97) = 21.42$ ,  $p < 0.001$ ,  $d = 0.94$  and GPA  $F(1, 97) = 4.12$ ,  $p = 0.045$ ,  $d = 0.41$  both remained significant covariates. Individuals that had higher ACT scores and GPAs tended to do better on the exam regardless of which condition they were randomly assigned to. A one-way ANCOVA was run on all question types compared to the instruction only control group for adjusted Dirkzwager scores. ACT score and GPA were left in the model to be used as covariates. A significant difference was found  $F(1, 118) = 8.81$ ,  $p = 0.004$ ,  $d = 0.55$  suggesting that testing in any format ( $M = 1273.67$ ,  $SD = 332.47$ ) tended to enhance learning more so than an instruction only ( $M = 1046.65$ ,  $SD = 339.54$ ) condition. ACT score  $F(1, 118) = 19.40$ ,  $p < 0.001$ ,  $d = 0.81$  and GPA  $F(1, 118) = 4.44$ ,  $p = 0.037$ ,  $d = 0.41$  continued to be significant covariates. Again, Individuals that had higher ACT scores and GPAs tended to do better on the exam regardless of which condition they were randomly assigned to.

## **DV: Metacognition**

A series of regressions were performed to see how well learners predicted their own performance, again this judgment of self performance task was given after the final statistical quiz was administered. For all regressions computed in this project, the predictor variable was the learner's predicted performance and the outcome variable was the learner's actual performance. For open ended question participants were able to significantly ( $F(1, 158) = 123.50, p < .001, d = 1.77$ ) predict ( $M = 67.47, SD = 20.23$ ) their actual performance ( $M = 13.18, SD = 3.07, \text{mean percentage} = 65.88$ ). Though over confident, learners were able to significantly ( $F(1, 158) = 48.99, p < .001, d = 1.11$ ) predict ( $M = 74.41, SD = 19.32$ ) their own performance ( $M = 14.43, SD = 2.67, \text{mean percentage} = 72.13$ ) on multiple choice questions. In terms of multiple evaluation questions, learners were also able to significantly ( $F(1, 158) = 108.19, p < .001, d = 1.67$ ) predict ( $M = 72.38, SD = 20.34$ ) their actual performance ( $M = 1427.07, SD = 254.09, \text{mean percentage} = 71.35$ ). Though participants were able to accurately predict their own performance they seem to have greater insight to that information when administered questions in an open ended or multiple evaluation type format, as indicated by the respective effect sizes.

A repeated measures ANOVA was conducted to see if there was a difference between difficulty levels of question type. A significant main effect was found  $F(1, 159) = 100.01, p < 0.001, d = 1.60$ . A Bonferroni multiple comparison suggests that the open ended questions ( $M = 3.88, SD = 1.40$ ) were significantly more difficult than the multiple evaluation and multiple choice type questions. Multiple evaluations ( $M = 3.26,$

$SD = 1.53$ ) were perceived to be significantly more difficult than multiple choice type ( $M = 2.99, SD = 1.27$ ) questions.

Chi-squares were conducted to see which learning conditions participants preferred and to see which they thought would result in the most amount of learning. Participants tended to prefer multiple choice (104) questions over open ended (17) questions, multiple evaluation type (34) questions, or no questions (5)  $\chi^2(3, N = 160) = 147.15, p < .001$ . Learners tended to think that open ended (68) or multiple choice (53) type questions would be more beneficial for learning than multiple evaluation (31) or no questions (8)  $\chi^2(3, N = 160) = 51.45, p < .001$ . Learner's preference for multiple choice and open ended questions will be discussed (see Table 4).

**Table 4. Participant's Preference and Best Predicted for Learning**

	Method selected			
	Open ended	Multiple choice	Multiple evaluation	No questions
Preference	17	104	34	5
Best for learning	68	53	31	8

### **Experiment 1 Discussion**

Across all dependent measures a significant effect for retrieval practice was found. The data suggests that as long as one is tested in some fashion during the learning phase of the experiment, pretest questions seem to be beneficial for learning statistical learning within an electronic environment. It should be noted that this effect was found for both multivariate and univariate data analyses. Our hypothesis that conditions with questions during the learning phase of the experiment would boost

learning is supported while our hypothesis that the three testing types may differ was rejected. It should be noted that retrieval practice in Experiment 1 was incorporated without the use of feedback. Results indicate that providing learners with questions (regardless of format) during an instruction phase seems to improve their long term performance about one letter grade.

Results for the segmentation type manipulation suggest that segmentation had no impact on statistical learning in an electronic environment. Segmentation does not help nor hinder learning within this study. It could be the case that the exam is too easy and thus learners do not need the information to be segmented. If the material was more complicated or difficult then the segmentation during instruction may then be beneficial for learning. Our hypothesis that segmentation would boost learning was not supported.

Metacognitive measures suggest that learners were able to predict their own performance regardless of test format. However, open ended and multiple evaluation formats tended to result in better predictions of the learners own performance as indicated by the amount a variance accounted for listed above. This is likely due to open ended and multiple evaluation question formats being more cognitively effortful than a multiple choice format. For example open ended questions elicit cued recall which is often known to be more difficult than recognition memory tasks such as a multiple choice test (Rowland, 2014). Multiple evaluation's on the other hand are a type of a recognition memory task as well, however it is likely that the added judgment of learning dimension of the task is making it cognitively more difficult. This difficulty sometimes behoove memory and learning in the long run (Maddox & Balota, 2015). Such difficulty, as perceived by the learners, tended to behoove metacognitive insight.

One way to improve learning even further while using retrieval practice is to provide feedback (Hayes et al., 2013). Thus, feedback will be added to Experiment 2 in an effort to further boost learning of the statistical information. Corrective feedback will be able to be utilized in all of the previous conditions administered within Experiment 1. It is hypothesized that feedback will improve learning even more so than the previous testing only conditions. Feedback is also expected to improve metacognitive insight into a learners own knowledge thereby letting them know directly what they do and do not know (Butler, Karpicke, & Roediger, 2008). Segmentation, though not beneficial during experiment 1 will continue to be investigated.

## **Experiment 2**

### **Participants**

Two-hundred and sixty-three students volunteered for Experiment 2. Fifty-two participants were excluded from the data analyses for failing to return to part two of the experiment. Of the remaining 211 participants, 152 were female with ages ( $M = 18.62$ ,  $SD = 1.02$ ) ranging from 18 to 25. Class level reported indicated 79.6% freshmen, 15.6% sophomores, 3.8% juniors, and 0.9% seniors. Participants reported an average GPA of ( $M = 3.45$ ,  $SD = 0.58$ ) and an ACT score of ( $M = 26.55$ ,  $SD = 4.07$ ). Additional descriptive statistics are listed in Table 7.

**Table 5. Demographics for Experiment 2**

	Count	Percentage
<b>Statistics Knowledge</b>		
I have not taken a class nor do I have any prior experience with formal statistics	138	65.4%
I have not taken a class but I do have some experience with formal statistics	10	4.7%
I have taken a class at the high school level	30	14.2%
I have taken a class at the college level	33	15.6%
I have taken a class at the graduate level	0	0%
<b>Research Methods Knowledge</b>		
I have not taken a class nor do I have any prior experience with research methods	131	62.1%
I have not taken a class but I do have some experience with research methods	59	28%
I have taken a class at the high school level	9	4.3%
I have taken a class at the college level	12	5.7%
I have taken a class at the graduate level	0	0%
<b>Math Knowledge</b>		
High school mathematics	39	18.5%
Some college mathematics	39	18.5%
College algebra	53	25.1%
College calculus	73	34.6%
College linear algebra or higher	7	3.3%

**Materials**

The majority of materials used in Experiment 2 were identical to those used in Experiment 1. A new set of 20 statistical questions was developed to assess knowledge transfer. Similar to the original 20 questions these questions also aimed to assess the previous topics of correlations, measurement, research design, and sampling. The primary investigator developed these questions to assess near or medium-near

knowledge transfer. The new transfer questions were also designed to be administered within open ended, multiple choice, or multiple evaluation type formats.

### **Procedure and design**

Like experiment 1, experiment 2 also utilized a 3 (question type) X 2 (segmentation type) factorial design with an added control group. However, because feedback is being used in experiment 2 dependent measures were changed to be between type design instead of within during the retention portion of the experiment. After answering the original 20 questions participants were asked a series of metacognitive questions and then were administered the 20 transfer questions (and then asked another series of metacognitive questions). Due to the design change a multivariate analysis is no longer appropriate so it will not be analyzed. Similar to the data analysis in question 1, each question type will be analyzed individually. Non feedback conditions from experiment 1 will also be compared to feedback conditions conducted within experiment 2.

Part 1 of Experiment 2 is the same as part 1 of experiment 1 except for the added corrective feedback. After each question was asked corrective feedback was administered. For the multiple choice questions participants were told whether or not they got the previous question correct, were shown the correct answer, and were then given a running total of how many answers they had answered correctly so far. For the multiple evaluation questions participants were told the correct answer as well as how much credit they received or had taken away for their respective answer. A running total was also given to the learners so that they could see clearly how they were doing

on the examination. Participants that received the questions in an open ended format were given the correct answer and asked to grade their own work. Possible grades that they could give to themselves were: "correct," "partially correct," and "incorrect." A running tally of how many answers were answered correctly were not provided for the open ended condition.

During part 2 of the experiment participants were administered the original 20 questions with immediate feedback provided after each question. The format of question type remained consistent between part 1 and part 2 (e.g. if the learner had multiple choice questions during part 1 of the study they would also have multiple choice questions during part 2). Metacognitive questions were then asked (how difficult was the previous set of questions and how well do you think you performed on the quiz). Next, the 20 transfer questions were asked with immediate corrective feedback provided after every answer. After the transfer questions were administered the same metacognitive questions were asked. At the end of the part 2 participants were also asked which question format they think they would like the most (after a brief explanation of the other conditions) and which question format administered during part 1 would result in the best overall learning. Upon completion participants were debriefed, thanked for their time, and given course credit.

The control condition in experiment 2 received no questions during part 1 of the study. However, during part 2 the control condition was administered questions in the multiple evaluation format only. All other aspects of part 2 of the control condition were identical to the other conditions. It is hypothesized that feedback will improve learning even more so than the previous testing only conditions. Feedback is also



expected to improve metacognitive insight into a learners own knowledge there by letting them know directly what they do and do not know. Segmentation is expected to be helpful for learning when feedback is added to the experimental design (rather than not segmenting).

## **Results (Experiment 2)**

### **DV: Open Ended**

The open ended questions were graded by 3 independently trained research assistants. Inter rater reliability statistics were computed for the original 20 questions ( $\alpha = 0.97$ ) as well as the 20 transfer questions ( $\alpha = 0.93$ ) used in experiment 2. A breakdown of each original question's inter rater reliability can be seen in table 6. A breakdown of each transfer question's inter rater reliability can be seen in table 7. Each item for the original 20 questions can be seen in Appendix 1 and all 20 transfer question items can be found in Appendix 2.

**Table 6. Experiment 2 Inter-rater Reliability**

<b>Question Number</b>	<b>Cronbach's <math>\alpha</math></b>
1	0.86
2	0.92
3	1
4	0.92
5	0.88
6	0.96
7	1
8	1
9	0.97
10	1

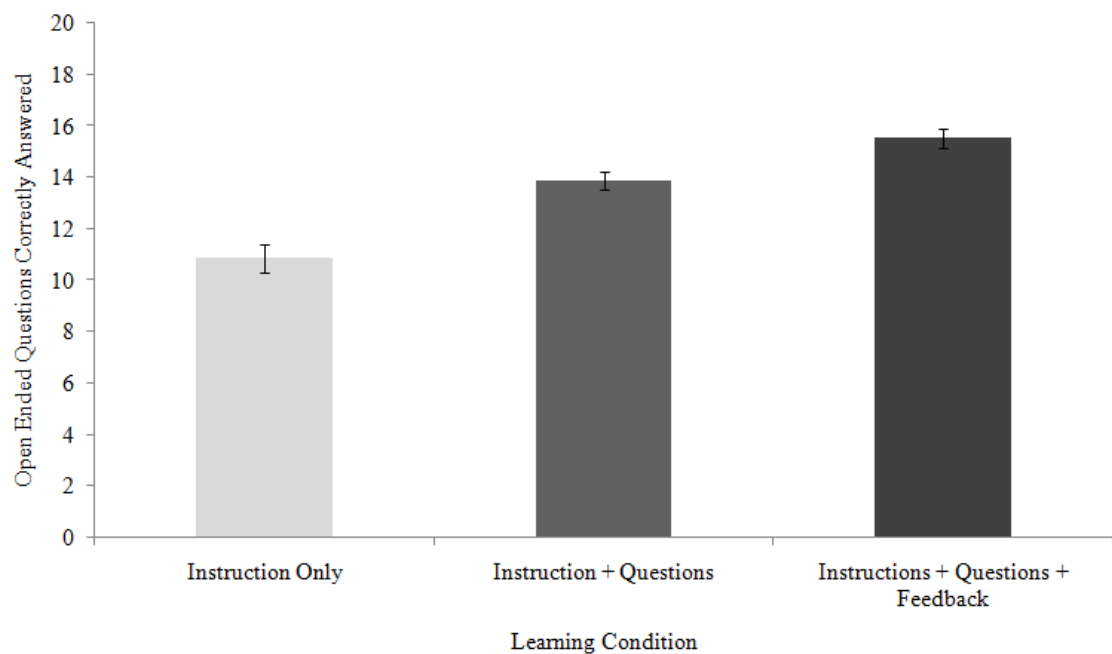
11	0.96
12	0.80
13	0.98
14	0.91
15	0.97
16	0.92
17	0.95
18	1
19	1
20	1
Total	0.97

**Table 7. Experiment 2 Transfer Inter-rater Reliability**

<b>Question Number</b>	<b>Cronbach's <math>\alpha</math></b>
1	0.87
2	1
3	0.89
4	0.95
5	0.94
6	0.92
7	0.99
8	1
9	1
10	0.90
11	0.98
12	0.61
13	0.95
14	0.90
15	1
16	0.73
17	0.99
18	0.40
19	0.99
20	0.94
Total	0.93

An initial ANCOVA suggested that ACT score, GPA, and Gender were the only significant covariates; all other covariates were dropped from the analysis. A (2 feedback type) X (2 segmentation type) ANCOVA with ACT score, GPA, and Gender was analyzed. No significant interaction was found  $F(1, 66) = 0.88$ , nor a main effect

of segmentation  $F(1, 66) = 0.35$ . However, a significant main effect for feedback type was found  $F(1, 66) = 10.72, p = 0.002, d = 0.81$  suggesting that providing feedback behooved learning ( $M = 15.53, SD = 2.62$ ) over not providing feedback ( $M = 13.87, SD = 2.49$ ) for open ended questions (see Figure 4). ACT score ( $F(1, 66) = 16.53, p < 0.001, d = 1.00$ ), GPA ( $F(1, 66) = 7.84, p = 0.007, d = 0.70$ ), and Gender ( $F(1, 66) = 6.92, p = 0.01, d = 0.67$ ) remained significant covariates. Males with higher ACT scores and higher GPAs tended to score higher on the statistical test regardless of which condition they were assigned to. The unexpected gender effect will be discussed within the discussion section of this experiment.



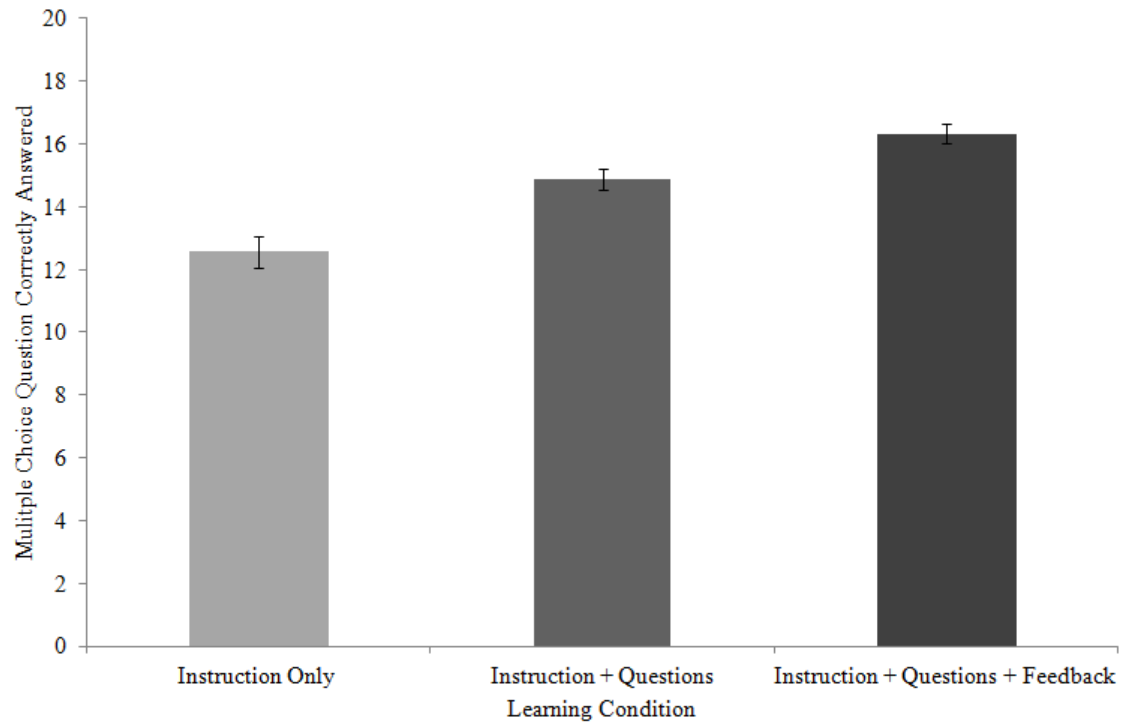
**Figure 4. Correctly Answered Open Ended Questions by Learning Condition**

A dependent t-test was conducted to see how learners performed on the original test compared to the transfer test. Learners in experiment 2 performed significantly worse  $t(51) = 14.30, p < 0.001$ , on the transfer test ( $M = 11.27, SD = 3.64$ ) than on the

original test ( $M = 15.78, SD = 2.69$ ) administered. Learners also perceived the transfer test ( $M = 4.51, SD = 1.46$ ) to be more difficult than the original test ( $M = 3.08, SD = 1.38$ ),  $t(51) = -8.49, p < 0.001$ .

### **DV: Multiple Choice**

An initial ANCOVA suggested that GPA and ACT score were the only significant covariates so all other covariates were dropped from the analysis. A (2 feedback type) X (2 segmentation type) ANCOVA was analyzed with GPA and ACT score as covariates. No interaction was found between feedback and segmentation type factors  $F(1, 65) = 1.80$ . No main effect was found for segmentation type  $F(1, 65) = 0.002$ . However, a main effect was found for feedback type  $F(1, 65) = 9.65, p = 0.003, d = 0.77$  suggesting that feedback in conjunction with testing ( $M = 16.32, SD = 2.29$ ) was more beneficial for long term retention than testing alone ( $M = 14.88, SD = 2.26$ ), see Figure 5. GPA ( $F(1, 65) = 6.32, p = .01, d = 0.63$ ) and ACT ( $F(1, 65) = 11.55, p = .001, d = 0.84$ ) score remained significant covariates. Our hypothesis that segmentation would be beneficial for learning was not supported while our hypothesis that feedback would be beneficial for learning was supported.



**Figure 5. Correctly Answered Multiple Choice Questions by Learning Condition**

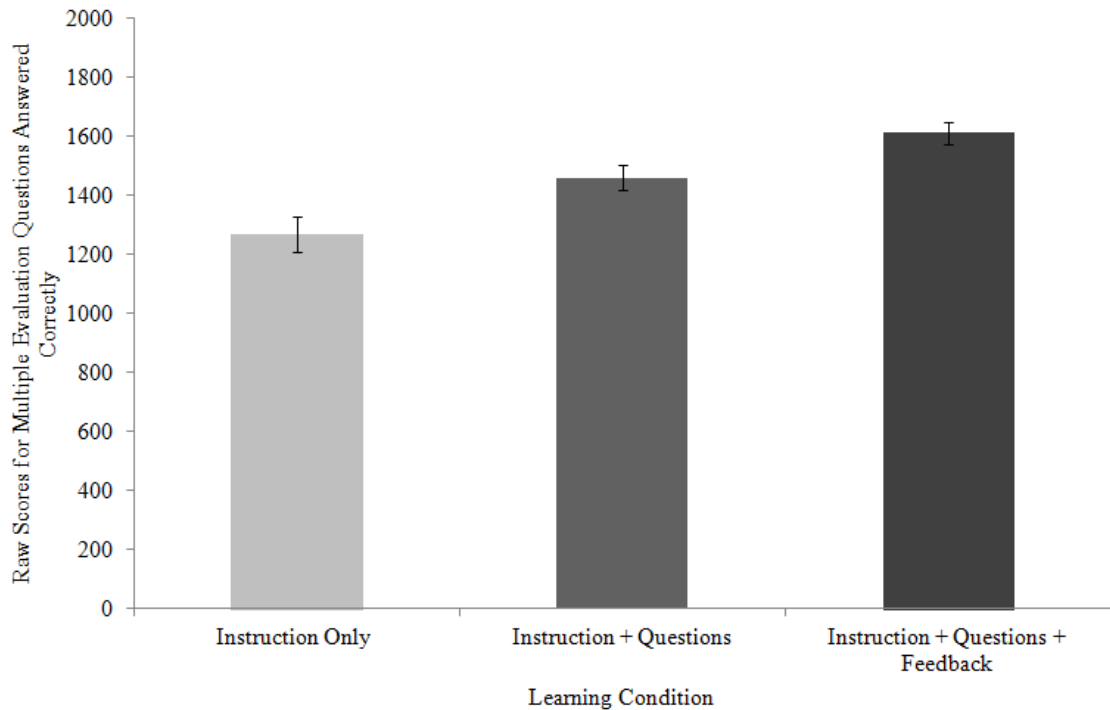
A dependent t-test was conducted to see if there was a difference between multiple choice type questions in experiment 2 and the transfer questions. Learners performed significantly ( $t(50) = 7.16, p < .001$ ) worse on the transfer question test ( $M = 13.71, SD = 3.50$ ) than on the original test ( $M = 16.67, SD = 2.07$ ). Learners also perceived the transfer test ( $M = 4.04, SD = 1.43$ ) as more difficult than the original test ( $M = 2.45, SD = 1.00$ );  $t(50) = -11.30, p < .001$ .

### **DV: Multiple Evaluations**

According to an initial ANCOVA statistical knowledge was the only significant covariate; all other covariates were dropped from the analysis. A (2 feedback type) X

(2 segmentation type) ANCOVA was analyzed. No interaction was found between segmentation type and feedback type  $F(1, 93) = 0.00$ . No main effect was found for segmentation type  $F(1, 93) = 0.20$ . However, a main effect of feedback type was found ( $F(1, 93) = 7.04, p = 0.009, d = 0.55$ ) favoring providing learners with feedback ( $M = 1612.95, SD = 285.19$ ) over not providing learners with feedback ( $M = 1460.11, SD = 286.56$ ). Prior statistics knowledge remained a significant covariate ( $F(1, 93) = 7.63, p = 0.007, d = 0.59$ ) suggesting that individuals with high prior statistical knowledge tended to do better on the final examination in a multiple evaluation type format regardless of what condition they were assigned to.

Again, for experiment 2 the multiple evaluation type format was the only format that had an added control condition in which learners received no questions during the instructional phase of the experiment. An additional one-way ANCOVA was conducted to see the impact of questions during the learning phase of the experiment vs having questions with feedback vs having no questions at all during the instructional phase of the experiment (prior statistical knowledge was kept in the model as a covariate). A main effect of condition type was found  $F(2, 117) = 12.23, p < 0.001, d = 0.91$ . A multiple comparison Bonferroni yielded that questions with feedback ( $M = 1612.72, SD = 288.71$ ) performed significantly better than questions alone ( $M = 1460.39, SD = 308.51$ ) or no questions ( $M = 1267.61, SD = 286.12$ ) during the learning phase of the experiment (see Figure 6). Prior knowledge of statistics remained a significant covariate  $F(1, 117) = 8.17, p = 0.005, d = 0.91$ . Our hypothesis that segmentation would be beneficial for learning was not supported, but our hypothesis that feedback would be beneficial for learning was supported.



**Figure 6. Correctly Answered Multiple Evaluation Questions by Learning Condition**

An additional one-way ANCOVA was run to see if question type had an impact on transfer question performance. An initial ANCOVA suggested that ACT score was a significant covariate so it was included within the model. The one-way ANCOVA revealed that there was no difference on the transfer question test between individuals that had questions with feedback vs individuals that had no questions during the initial learning phase of the experiment  $F(1, 67) = 1.19$ . ACT score remained a significant covariate within the model  $F(1, 67) = 21.14, p < 0.001, d = 1.12$ . A dependent t-test suggests that the transfer questions ( $M = 1277.39, SD = 295.09$ ) were significantly more difficult than the original 20 questions ( $M = 1512.54, SD = 366.28$ );  $t(78) = 7.08, p < .001$ .

A (2 feedback type) X (2 segmentation type) ANCOVA was analyzed using Dirkzwager adjusted scores. Prior statistical knowledge was left in the model as a covariate. No interaction was found between feedback and segmentation type conditions  $F(1, 93) = 0.97$ . No main effect was found for segmentation type  $F(1, 93) = 0.19$ , however a main effect was found for the feedback factor  $F(1, 93) = 7.30, p = 0.008, d = 0.55$ . The data trend suggests that immediate feedback ( $M = 1485.15, SD = 378.21$ ) tended to boost learning more so than not providing feedback ( $M = 1279.94, SD = 376.30$ ). Prior statistical knowledge remained a significant covariate within the model  $F(1, 93) = 7.69, p = 0.007, d = 0.59$ , suggesting that individuals with higher levels of statistical knowledge performed better on the test regardless of what group they were randomly assigned.

An additional one-way ANCOVA was conducted to see the impact of questions during the learning phase of the experiment vs having questions with feedback vs having no questions at all during the instructional phase of the experiment for adjusted Dirkzwager scores (prior statistical knowledge was kept in the model as a covariate). A main effect was found between the three learning conditions  $F(2, 117) = 12.35, p < 0.001, d = 0.91$ . A Bonferroni multiple comparison suggested that questions that were provided with immediate feedback ( $M = 1484.58, SD = 372.81$ ) were more beneficial for learning than the question condition without feedback ( $M = 1280.44, SD = 379.82$ ) and the instruction only ( $M = 1024.53, SD = 379.83$ ) manipulation. A significant difference was found between the question and instruction only conditions. Prior statistical knowledge remained a significant covariate  $F(1, 117) = 8.25, p = 0.005, d = 0.55$  suggesting that individuals with higher prior statistical knowledge tended to



perform better on the exam than individuals with no or lower prior knowledge. In general the adjusted dirkzwager scores tended to follow the same pattern of results that were found with the percentage of correct raw multiple evaluation scores that were computed above.

### **DV: Metacognition**

A series of regression were conducted to see how proficient participants were at predicting their own performance. Again, this judgment of performance task was done after learners had completed the respective test. In this particular experiment the learners should be performing very well because they are being provided feedback after every answer. Recall that multiple choice and multiple evaluation type questions were provided with running feedback for total progress (e.g. you have answered 13 out of 15 correct or your total score is 1250). For the original 20 questions participants were able to significantly predict ( $M = 84.61$ ,  $SD = 13.41$ ) their own performance ( $M = 16.67$ ,  $SD = 2.04$ , mean percentage = 83.35) within a multiple choice format  $F(1, 49) = 64.22$ ,  $p < .001$ ,  $d = 2.29$ . For the 20 transfer questions learners were able to significantly predict ( $M = 68.82$ ,  $SD = 18.86$ ) their own performance ( $M = 13.71$ ,  $SD = 3.49$ , mean percentage = 68.55) within a multiple choice format  $F(1, 49) = 87.39$ ,  $p < .001$ ,  $d = 2.67$ . Learners were able to predict ( $M = 80.00$ ,  $SD = 20.47$ ) their own performance ( $M = 1614.07$ ,  $SD = 338.01$ , mean percentage = 80.70) within a multiple evaluation format on the original 20 question test  $F(1, 54) = 94.11$ ,  $p < .001$ ,  $d = 2.67$ . Participants were also able to significantly predict ( $M = 71.34$ ,  $SD = 18.94$ ) their own performance ( $M =$

1275.32,  $SD = 285.17$ ) within a multiple evaluation format for the set for 20 transfer questions  $F(1, 54) = 18.04, p < .001, d = 1.15$ . For the original 20 questions participants were able to significantly predict ( $M = 79.62, SD = 20.62$ ) their own performance ( $M = 15.71, SD = 2.67, \text{mean percentage} = 78.55$ ) within an open ended format  $F(1, 49) = 20.83, p < .001, d = 1.31$ . For the 20 transfer questions learners were able to significantly predict ( $M = 62.89, SD = 24.64$ ) their own performance ( $M = 11.15, SD = 3.63$ ) within an open ended format  $F(1, 49) = 31.48, p < .001, d = 1.64$ .

A series of dependent t-tests were conducted to see if transfer questions were also perceived to be more difficult than the original questions. Within a multiple choice format participants perceived the transfer test ( $M = 4.04, SD = 1.43$ ) to be more difficult than the original test ( $M = 2.45, SD = 1.00$ ),  $t(50) = -11.30, p < .001$ . When administered in a multiple evaluation format the transfer test ( $M = 3.73, SD = 1.35$ ) was also perceived to be more difficult than the original test ( $M = 2.91, SD = 1.27$ ) administered,  $t(55) = -5.09, p < .001$ . The transfer test ( $M = 4.51, SD = 1.36$ ) was also perceived to be more difficult than the original set ( $M = 3.08, SD = 1.36$ ) of questions in an open ended format  $t(50) = -8.48, p < .001$ . Regardless of test format the transfer test was perceived to be more difficult by learners.

A series of chi-squared tests were conducted to see which learning condition learners preferred and to see which condition they thought would result in greatest boost in learning. Participants tended to favor multiple choice questions (143) over open ended questions (11), multiple evaluation type questions (18), or no questions (8)  $\chi^2(3, N = 180) = 285.73, p < .001$ . Participants also tended to think that the multiple choice question format (87) would result in the greatest boost in learning over open ended (47),

multiple evaluation (43), or no questions (3)  $\chi^2 (3, N = 180) = 78.58, p < .001$ .

Learner's preference for multiple choice questions will be discussed (see Table 10).

**Table 8. Experiment 2 Participant's Preference and Best Predicted for Learning**

	Method selected			
	Open ended	Multiple choice	Multiple evaluation	No questions
Preference	11	143	18	8
Best for learning	47	87	43	3

## Experiment 2 Discussion

Across all testing type formats immediate corrective feedback was found to be beneficial for learning. The addition of corrective feedback to the learning procedure typically boosted learner's performance about a letter grade. Taking experiment 1's findings into consideration, the no test condition can be thought of as a base line control; while testing alone tended to boost learner's performance a letter grade, providing learners with immediate corrective feedback tended to boost participant's abilities yet another letter grade (see Figure 6). The trend of the data suggests that providing learner's with tests with immediate corrective feedback boosts the learning by 2 letter grades over the instruction only control. For a breakdown of average performance of each manipulation by testing format see Table 9. Our hypothesis that immediate feedback would be more beneficial for learning than testing alone was supported.

**Table 9. Experiment 1 & 2 Learning Condition Performance by Final Test Format**

	Final Test Format		
	Multiple Choice	Open Ended	Multiple Evaluation
Instruction Only	12.83 (0.55) D	10.85 (0.57) F	1267.61 (59.66) D
Instruction + Testing	14.88 (0.33) C	13.87 (0.36) D	1460.39 (43.63) C
Instruction + Testing + Feedback	16.32 (0.32) B	15.53 (0.36) C	1612.72 (38.58) B

\*The dependent variables listed are means in the respective test format's units. Standard errors are in parentheses.

Across all testing formats segmentation was found to neither help nor hinder learning. Similar to experiment 1, the addition to feedback had no impact on the effectiveness of segmentation. Again, this could be because the test administered was too easy. Had the exam been more difficult segmentation may have behooved learning. The data suggests that our hypothesis that segmentation would benefit learning was not supported.

When feedback was administered, all learners were able to significantly predict their own performance regardless of test format. For the majority of test formats the addition of feedback aided learners metacognitive ability to judge their own performance. It should be noted again that in the multiple choice and multiple evaluation conditions learners were also given a running tally of their overall performance. Due to the difficulty of grading open ended questions as soon as they were submitted learners were only given the immediate correct answer upon each answer submission. In experiment 2 retrieval practice with feedback tended to behoove metamemory knowledge particularly for the multiple choice and multiple evaluation

conditions. A running tally of how well the learners were performing tended to benefit metacognitive understanding of their own knowledge.

In all testing formats participants performed significantly worse on the transfer test than they did on the set of original test questions administered during the learning phase of the experiment. This was expected to happen; the transfer of newly learned information is difficult. However, the finding that both the instruction only and testing with immediate feedback learning conditions performed at the same level was surprising. In other words, on the transfer test no difference was seen between the instruction only and questions with feedback in terms of performance level within a multiple evaluation format. Originally the transfer test was intended to measure near transfer knowledge, however, it could be the case that the transfer test is instead measuring medium to far levels of transfer. Far transfer questions are typically found to be more difficult than near transfer questions, due to the present transfer test possibly containing far transfer questions this could be the reason for our findings.

Experiment 3 within this project will continue to investigate the finding that the benefits of retrieval practice do not transfer beyond knowledge of the original test. More specifically, experiment 3 will investigate whether or not the testing effect (without feedback) will have a positive benefit for learner's transfer knowledge. Previous research suggests that the benefits associated with retrieval practice should indeed transfer (Butler, 2010). It is hypothesized that retrieval practice alone should be beneficial for learner's transfer of new knowledge to novel situations. Because little is known about segmentation and its impact on transfer knowledge, the segmentation phenomenon will continue to be inquired.

## Experiment 3

### Participants

One-hundred and four students volunteered for Experiment 3. Twenty participants were excluded from the data analyses for failing to return to part two of the experiment. Of the remaining 84 participants, 47 were female with ages ( $M = 19.34$ ,  $SD = 2.54$ ) ranging from 18 to 41. Class level reported indicated 73.8% freshmen, 19% sophomores, 6% juniors, and 1.2% seniors. Participants reported an average GPA of ( $M = 3.34$ ,  $SD = 0.47$ ) and an ACT score of ( $M = 26.30$ ,  $SD = 3.49$ ). Additional descriptive statistics are listed in Table 10.

**Table 10. Demographics for Experiment 3**

	Count	Percentage
<b>Statistics Knowledge</b>		
I have not taken a class nor do I have any prior experience with formal statistics	44	52.4%
I have not taken a class but I do have some experience with formal statistics	7	8.3%
I have taken a class at the high school level	16	19%
I have taken a class at the college level	17	20.2%
I have taken a class at the graduate level	0	0%
<b>Research Methods Knowledge</b>		
I have not taken a class nor do I have any prior experience with research methods	56	66.7%
I have not taken a class but I do have some experience with research methods	23	27.4%
I have taken a class at the high school level	4	4.8%
I have taken a class at the college level	1	1.2%
I have taken a class at the graduate level	0	0%
<b>Math Knowledge</b>		

High school mathematics	1	1.2%
Some college mathematics	18	21.4%
College algebra	18	21.4%
College calculus	44	52.4%
College linear algebra or higher	3	3.6%

---

## **Materials**

Materials used in Experiment 3 were essentially identical to those used in Experiment 2. No corrective feedback was provided for learners within this experiment. Transfer questions from experiment 2 were included.

## **Procedure and Design**

Part 1 of Experiment 3 was identical to Experiment 1 except there was no control group included. Part 2 of the Experiment 3 was identical to part 2 within Experiment 2 except there was no feedback. The goal of this experiment was to see how individuals performed on the transfer test when only given questions during the initial learning phase of the experiment (no feedback was given). The same patterns of results found in Experiment 1 are expected. Questions administered during the learning phase are hypothesized to behoove learning of that respective information. Improved learning is also expected to result in better performance on the transfer question test than individuals that were not given questions during the Instructional phase of the experiment. Though segmentation has seen limited to no success within this project it is

still hypothesized that segmented conditions will tend to perform better on the final statistical test than the non-segmented conditions.

### Results (Experiment 3)

#### DV: Open Ended

The open ended questions were graded by 3 independently trained research assistances. Inter rater reliability statistics were computed for the original 20 questions ( $\alpha = 0.96$ ) as well as the 20 transfer questions ( $\alpha = 0.92$ ) used in experiment 2. A breakdown of each original question's inter-rater reliability can be seen in table 11. A breakdown of each transfer question's inter-rater reliability can be seen in table 12.

**Table 11. Experiment 3 Inter-rater Reliability**

<u>Question Number</u>	<u>Cronbach's <math>\alpha</math></u>
1	1
2	1
3	1
4	0.91
5	1
6	0.95
7	1
8	1
9	0.99
10	1
11	0.97
12	0.51
13	0.95
14	1
15	0.93
16	0.78
17	0.95
18	0.98



19	0.97
20	0.98
Total	0.96

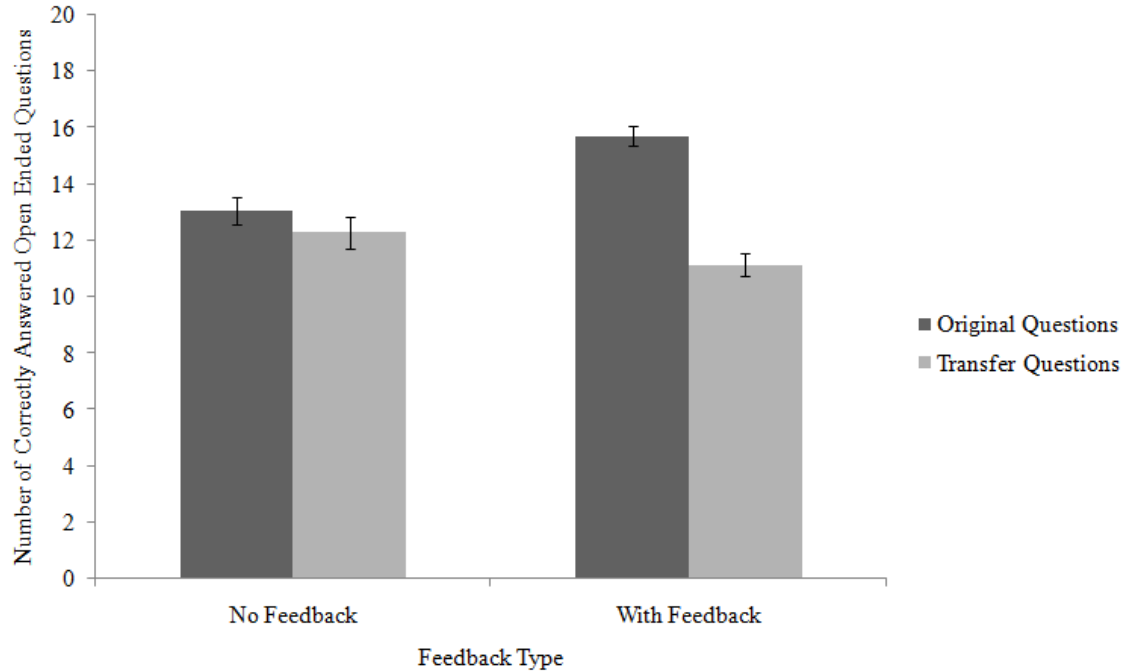
**Table 12. Experiment 3 Transfer Inter-rater Reliability**

Question Number	Cronbach's $\alpha$
1	0.70
2	0.95
3	0.97
4	0.92
5	0.92
6	0.82
7	0.90
8	0.94
9	0.97
10	0.88
11	0.96
12	0.58
13	0.97
14	0.81
15	0.93
16	0.77
17	1
18	0.68
19	0.94
20	0.83
Total	0.92

A (2 feedback type) X (2 segmentation type) ANCOVA was analyzed for transfer questions in an open ended format. A primary analysis suggests that ACT score and gender were the only significant covariate so those covariates were the only ones that were used in the current model. No interaction was found between feedback and segmentation type factors  $F(1, 70) = 0.97$ . No main effects were found for feedback ( $F(1, 70) = 1.63$ ) or segmentation ( $F(1, 70) = 2.67$ ) type manipulations for transfer questions within an open ended format. ACT score ( $F(1, 70) = 22.98, p < 0.001, d = 1.15$ ) and gender  $F(1, 70) = 6.25, p = 0.02, d = 0.59$  remained significant covariates.

Males with high ACT scores tended to perform well regardless of which condition they were randomly assigned to. The unexpected gender difference will be explained in the discussion section of experiment 3.

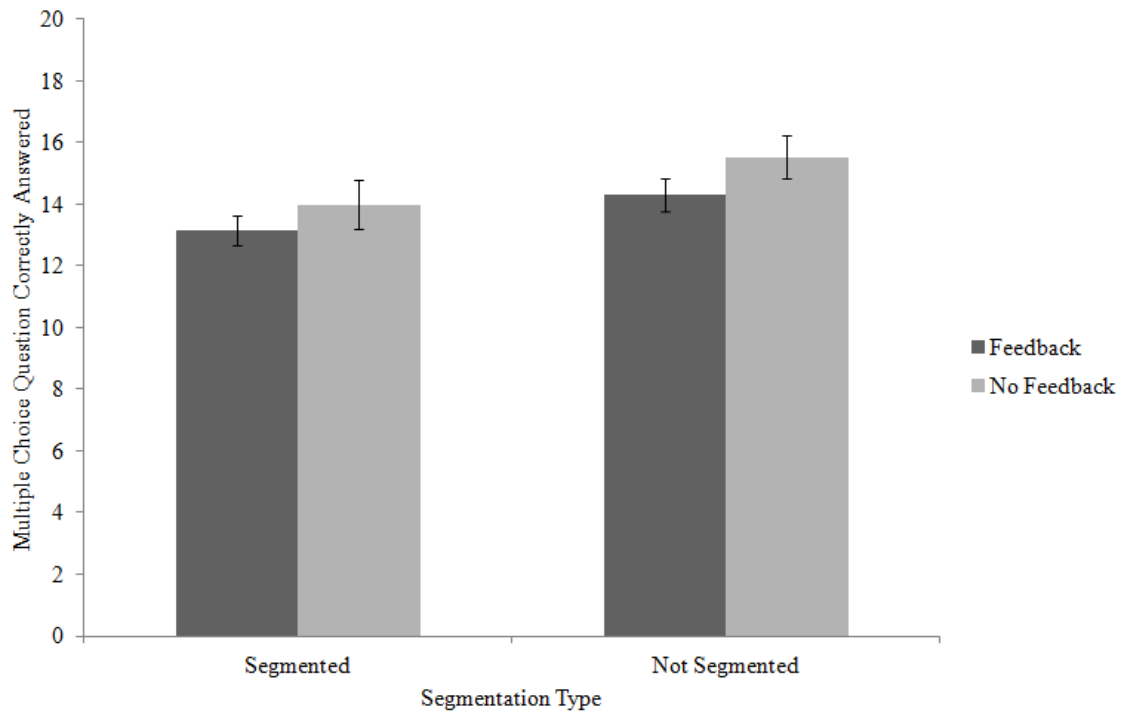
An initial analysis suggested that ACT score and gender were the only significant covariates; thus all other covariates were dropped from the analysis. A (2 feedback type) X (2 test type: original vs transfer) repeated measures ANCOVA was conducted with ACT score and gender as covariates. Feedback data was taken from experiment 2. A significant interaction was found between feedback and test type factors ( $F(1, 72) = 56.65, p < 0.001, d = 1.77$ ) such that feedback is aiding original questions but adding no benefit to transfer questions (see Figure 22). No main effect was found between feedback types  $F(1, 72) = 1.65$ . ACT score ( $F(1, 72) = 27.86, p < 0.001, d = 1.25$ ) and gender ( $F(1, 72) = 5.57, p = .02, d = 0.55$ ) remained significant covariates. Males with high ACT scores tended to score high on the statistical test regardless of which condition they were assigned.



**Figure 7. Correctly Answered Open Ended Questions by Feedback and Exam Type**

### **DV: Multiple Choice**

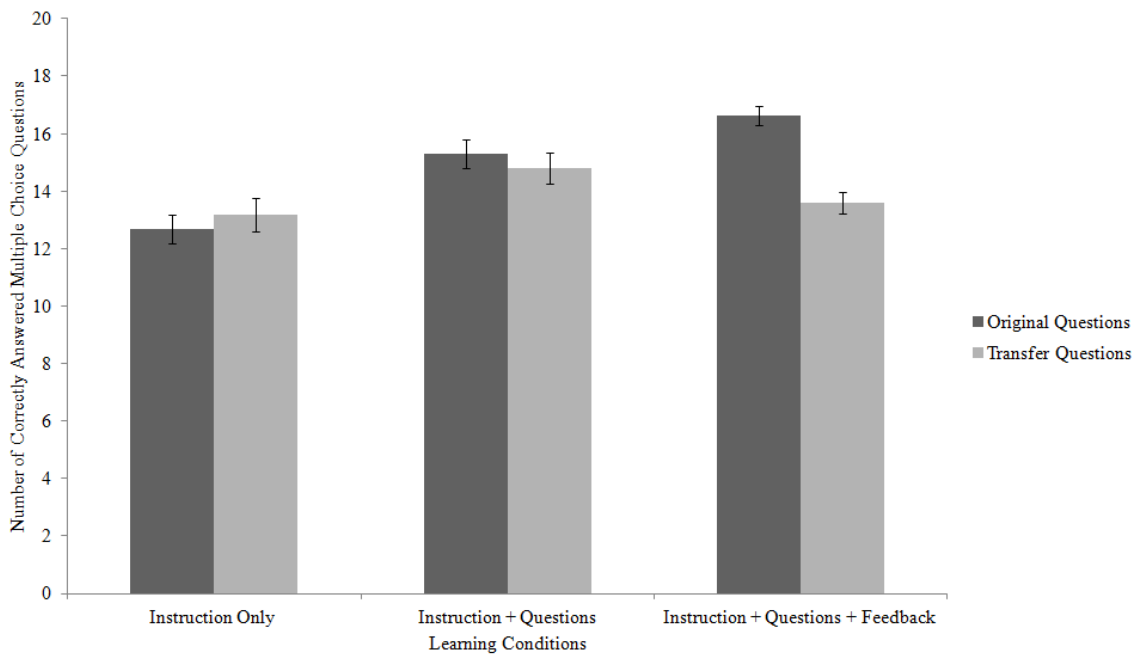
A (2 feedback type) X (2 segmentation type) ANCOVA was analyzed for transfer questions. An initial analysis suggested that statistical knowledge and ACT score were the only significant covariates. All other covariates were dropped from the analysis. No interaction was seen between feedback and segmentation factors  $F(1, 66) = 0.10$ . No main effect was found for feedback type  $F(1, 66) = 2.58$ , however a main effect was found for segmentation type  $F(1, 66) = 4.42, p = 0.04, d = 0.51$ . Learners in the segmentation type condition ( $M = 13.57, SD = 2.77$ ) performed significantly worse than conditions that were not segmented ( $M = 14.92, SD = 2.85$ ), see Figure 8. ACT scores ( $F(1, 66) = 25.29, p < 0.001, d = 1.25$ ) and statistical knowledge ( $F(1, 66) = 8.77, p = 0.004, d = 0.74$ ) remained significant covariates.



**Figure 8. Correctly Answered Multiple Choice Questions by Feedback and Segmentation Type**

A repeated measures ANCOVA was conducted on original and transfer questions in a multiple choice format. The Between factor levels consisted of 3 learning conditions: no questions (experiment 1 control), questions with feedback (experiment 2), and questions without feedback. It is worth noting that the no question control for transfer questions was originally in a multiple evaluation format but the scores were transformed so that they could be compared here. The previously significant covariates of ACT score and statistical knowledge were also left within the model. A significant interaction was found between the two dependent measures and the learning conditions  $F(2, 88) = 15.96, p < 0.001, d = 1.22$ . For the original questions administered, questions ( $M = 15.28, SD = 3.36$ ) and questions with feedback ( $M = 16.62, SD = 2.36$ ) performed better than the no questions ( $M = 12.67, SD = 2.45$ ) condition. However, for

transfer questions administered, the questions only condition ( $M = 14.80, SD = 2.86$ ) outperformed the no question ( $M = 13.13, SD = 2.73$ ) and questions with feedback ( $M = 13.60, SD = 2.77$ ) conditions (see Figure 9). It was found that the transfer questions ( $M = 13.84, SD = 2.74$ ) in general were more difficult than the original questions ( $M = 14.86, SD = 2.36$ )  $F(1, 88) = 13.24, p < 0.001, d = 0.13$ . ACT score ( $F(1, 88) = 30.84, p < 0.001, d = 1.19$ ) and prior statistical knowledge ( $F(1, 88) = 8.81, p = 0.007, d = 0.59$ ) remained significant covariates.

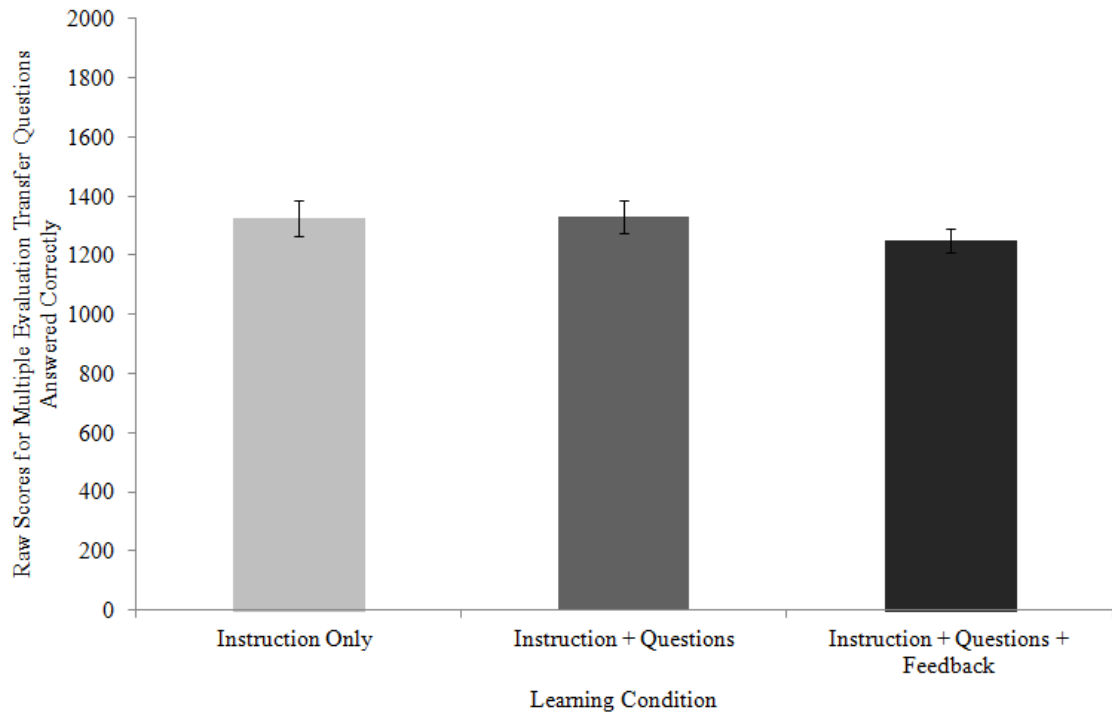


**Figure 9. Correctly Answered Multiple Choice Questions by Learning Condition and Exam Type**

## **DV: Multiple Evaluations**

A (2 feedback type) X (2 segmentation type) ANCOVA was analyzed for transfer questions. An initial ANCOVA analysis suggested that ACT score only was a significant covariate so all other covariates were dropped from the analysis. No interaction was found between the feedback and segmentation type factors  $F(1, 70) = 2.68$ . No main effect of feedback type was found  $F(1, 70) = 1.51$ , nor was there a main effect for segmentation type  $F(1, 70) = 3.12$ . ACT score remained a significant covariate  $F(1, 70) = 22.50, p < .001, d = 1.12$ . The data suggests that individuals with high ACT scores tended to have higher scores on the transfer test regardless of which experimental condition they were placed.

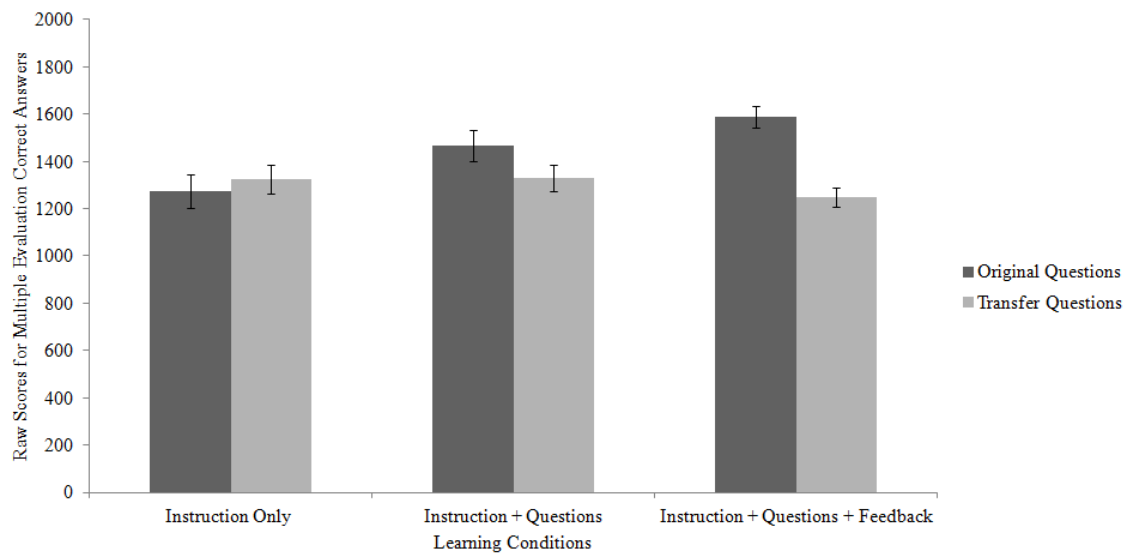
A one-way ANCOVA was conducted on transfer questions in the multiple evaluation type format. The levels consisted of 3 learning conditions: no questions (experiment 2 control), questions with feedback (experiment 2), and questions without feedback. An initial analysis suggested that ACT score was the only significant covariate so all other covariates were left out of the current model. No differences were found between the 3 learning conditions for transfer questions  $F(2, 92) = 0.95$  (see Figure 10). ACT score remained a significant covariate  $F(1, 92) = 27.49, p < .001, d = 1.09$ . The data trend suggests that individuals with high ACT scores tended to have higher scores on the transfer test regardless of which experimental condition they were placed.



**Figure 10. Correctly Answered Multiple Evaluation Transfer Questions by Learning Condition**

An additional repeated measures ANCOVA was conducted on the three learning conditions (no questions, questions with feedback, and questions without feedback) and the dependent measures (original questions vs transfer questions). A significant interaction was found between the dependent measures and the three learning conditions  $F(2, 92) = 22.60, p < .001, d = 0.33$ . For the initial 20 questions the question conditions ( $M = 1467.36, SD = 338.74$ ) and question with feedback ( $M = 1587.40, SD = 355.08$ ) conditions outperformed the no question ( $M = 1274.00, SD = 349.04$ ) condition. However, this effect was nullified when participants were administered the transfer questions (see figure 11). Transfer questions ( $M = 1300.39, SD = 292.97$ ) were found to be significantly more difficult than the original questions ( $M = 1442.92, SD = 348.81$ ) within a multiple evaluation format  $F(1, 92) = 11.27, p < .001, d = 0.70$ . ACT score

remained a significant covariate  $F(1, 92) = 17.02, p < .001, d = 0.87$ . Again, the data trend suggests that individuals with higher ACT scores tended to perform better on the dependent measures regardless of which condition they were randomly assigned to.



**Figure 11. Correctly Answered Multiple Evaluation Questions by Learning Condition and Exam Type**

A (2 feedback type) X (2 segmentation type) ANCOVA was analyzed for transfer questions in terms of Dirkwager adjusted scores. ACT score was left in the model as a covariate. No interaction was found between the feedback and segmentation type  $F(1, 70) = 1.47$ . No main effect of feedback type was found  $F(1, 70) = 0.53$ , nor was there a main effect for segmentation type  $F(1, 70) = 3.09$ . ACT score remained a significant covariate  $F(1, 70) = 22.50, p < .001, d = 1.06$ . The data suggests that individuals with high ACT scores tended to have higher scores on the transfer test regardless of which experimental condition they were administered. A one-way ANCOVA was conducted on transfer questions in the multiple evaluation type format



in terms of adjusted Dirkzwager scores. The levels consisted of 3 learning conditions: no questions (experiment 2 control), questions with feedback (experiment 2), and questions without feedback. An initial analysis suggested that ACT score was the only significant covariate so all other covariates were left out of the current model. No main effect was found for the learning condition type factor  $F(2, 92) = 0.61$ . In other words, no significant differences were seen for the adjusted Dirkzwager scores for the transfer questions when administered instruction only, instruction plus questions, or instruction plus questions with immediate feedback. ACT score remained a significant covariate within the model  $F(2, 92) = 26.26, p < .001, d = 1.06$ , suggesting that individuals that reported higher ACT scores tended to do better on the transfer test regardless of what condition the respective learner was assigned.

### **DV: Metacognition**

A series of regression were computed to see how well participants predicted their own performance. Again, this judgment of performance task was done after learners had completed the respective test was administered. For the original questions learners were able to predict ( $M = 72.00, SD = 23.47$ ) their own performance ( $M = 13.10, SD = 3.13$ , mean percentage = 65.50) significantly in an open ended format  $F(1, 28) = 51.04, p < .001, d = 2.67$ . Learners were also able to significantly predict ( $M = 64.00, SD = 25.24$ ) their own performance ( $M = 12.57, SD = 3.33$ , mean percentage = 62.85) for transfer question within an open ended format  $F(1, 28) = 64.00, p < .001, d = 2.98$ . For original questions learners were unable to significantly predict ( $M = 75.74,$

$SD = 15.95$ ) their own performance ( $M = 15.41$ ,  $SD = 2.69$ , mean percentage = 77.05) within a multiple choice format  $F(1, 25) = 2.95$ . However, for transfer questions participants were able to significantly predict ( $M = 71.11$ ,  $SD = 16.54$ ) their own performance ( $M = 14.93$ ,  $SD = 2.54$ , mean percentage = 74.65) within a multiple choice type format  $F(1, 25) = 5.03$ ,  $p = .03$ ,  $d = 0.90$ . In terms of multiple evaluations, learners were able to significantly predict ( $M = 70.00$ ,  $SD = 22.49$ ) their own performance ( $M = 1494.19$ ,  $SD = 335.81$ , mean percentage = 73.21) for the original set of questions  $F(1, 25) = 41.80$ ,  $p < .001$ ,  $d = 2.59$ . Participants were also able to significantly prediction ( $M = 65.74$ ,  $SD = 22.73$ ) their own performance ( $M = 1330.11$ ,  $SD = 347.91$ , mean percentage = 66.51) for the transfer questions within a multiple evaluation format  $F(1, 25) = 25.78$ ,  $p < .001$ ,  $d = 2.03$ .

A series of dependent t-tests were conducted to see if transfer questions were also perceived to be more difficult than the original questions. In an open ended format learners perceived transfer questions ( $M = 4.30$ ,  $SD = 1.42$ ) to be significantly more difficult than the original questions ( $M = 3.47$ ,  $SD = 1.42$ ),  $t(29) = -5.77$ ,  $p < .001$ . In a multiple choice format participants perceived transfer questions ( $M = 3.89$ ,  $SD = 1.40$ ) to be significantly more difficult than the original questions ( $M = 3.26$ ,  $SD = 1.30$ ),  $t(26) = -3.90$ ,  $p = .001$ . In a multiple evaluation format learners perceived transfer questions ( $M = 4.11$ ,  $SD = 1.40$ ) to be significantly more difficult than the original set of questions ( $M = 3.59$ ,  $SD = 1.35$ )  $t(26) = -2.56$ ,  $p = .02$ . Regardless of the format it would seem that the transfer test was perceived as more difficult than the original test administered to the learners.

A series of chi-squared tests were computed to see which learning conditions were preferred and which condition the learners thought would be more beneficial for learning. Participants tended to favor multiple choice questions (66) over open ended (10), and multiple evaluation type (8) formats  $\chi^2 (2, N = 84) = 77.43, p < .001$ . However, learners tended to think that multiple choice (37) and open ended (32) type formats would be more beneficial for learning than multiple evaluation (15) type formats  $\chi^2 (2, N = 84) = 9.50, p = .009$ . Learner's preferences for multiple choice and open ended type questions will be discussed (see Table 13).

**Table 13. Experiment 3 Participant's Preference and Best Predicted for Learning**

	Method selected			
	Open ended	Multiple choice	Multiple evaluation	No questions
Preference	10	66	8	0
Best for learning	32	37	15	0

### Experiment 3 Discussion

Across all conditions retrieval practice and retrieval practice with feedback is beneficial for the learning of statistical information within an online electronic learning environment. However, these findings seem to not be transferring to another test that is similar but different. Only in one condition within a multiple choice format did the benefits of retrieval practice seem to transfer. In all other conditions the benefits associated with retrieval practice and retrieval practice with feedback were not observed. This finding could be occurring for a few different reasons. It is a common finding that transfer is difficult (Campbell & Robert, 2008; Zhou et al., 2013) and it

could be the case that the transfer exam constructed was either too hard to notice mnemonic benefits or it could be the case that the transfer exam constructed consisted of questions that were far transfer rather than intended near transfer questions. Had the transfer test been better constructed it would likely have illuminated the benefits of testing and testing with feedback as mechanism for learning. Another possibility could be that the transfer exam is not measuring what we are attempting to measure. The validity of the transfer exam will be explored within the Item Response Theory (IRT) portion of this project below.

For the most part segmentation had no impact on learning transfer; however within a multiple choice format segmentation seemed to hinder knowledge transfer. Though this finding was small (about half a letter grade), it questions the robustness of the segmentation effect. Every manipulation within this project that has inquired into the segmentation effect has resulted in no added memory benefit. Another interesting finding was the gender effect within an open ended format. In general, within an open ended format males tended to outperform females regardless of which condition they were randomly assigned to. It should be noted that there was not an equal number of males and females (~70%) sampled. Of the males ( $M = 27.78$ ,  $SD = 5.41$ ) sampled they tended to report a higher ACT score than the females ( $M = 26.15$ ,  $SD = 3.43$ ), thus it could be the case that males sampled for this portion of the experiment had greater prior knowledge than the females sampled.

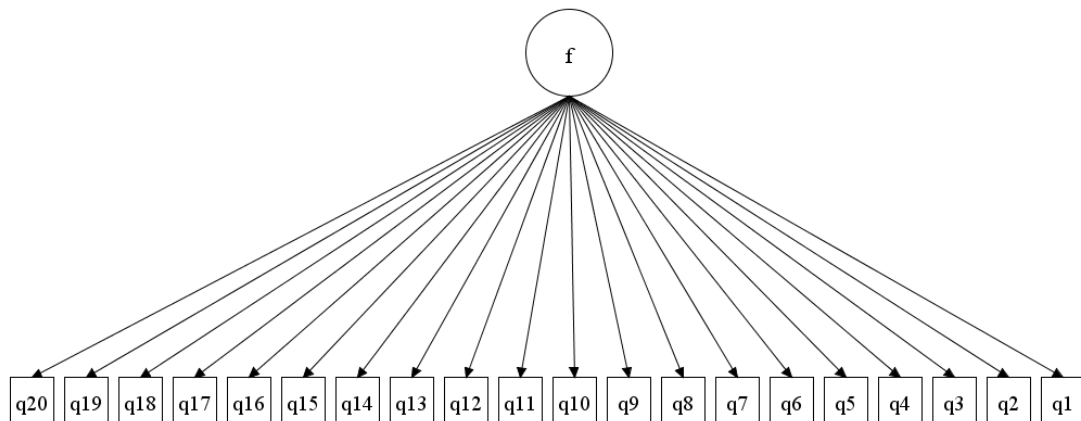
Similar to experiment 1 individuals that were randomly assigned to open ended or multiple evaluation question type formats tended to have greater insight into their own metacognitive knowledge than individuals assigned to the multiple choice format

manipulation. Again this finding is likely due to both the open ended and multiple evaluation formats being more cognitively effortful than the multiple choice type format. In addition, it is likely that the added judgment of learning task built into the multiple evaluation manipulation boosted learners' metamemory knowledge for that respective material being tested. The added metacognitive insight should be thought of as a benefit for multiple evaluation and open ended questions over the typical multiple choice type format. Furthermore if learners know what they know and what they do not know more accurately, students may then allocate time appropriately during study (Dunlosky et al., 2005).

In the next section of this project we will explore the quality of the tests used within this project. A basic Item Response Theory Model will be used to examine the validity and difficulty of both the original set of questions and the transfer question sets. Because Item Response Theory is not widely used by all researchers the next session will begin with a brief review of how the technique can be applied and utilized to make any test or measurement device more proficient at whatever it is that respective tool is trying to measure.

## Chapter 4: Item Response Theory

Item Response Theory (IRT) or Item Factor Analyses is a data analyses technique that determines qualities for measurements used on tests. One of the fundamental advantages that IRT provides psychometricians is a validity measure for latent traits. IRT often refers to this measure of validity as item discrimination or how well the respective question reflects what it is attempting to measure. For example, in this project we have been attempting to measure statistical knowledge that was taught during the learning phase of the experiment. The basics of IRT can tell psychometricians which items are easy or hard and which items are valid. Both the original and transfer test sets can be theoretically visualized in Figure 12.

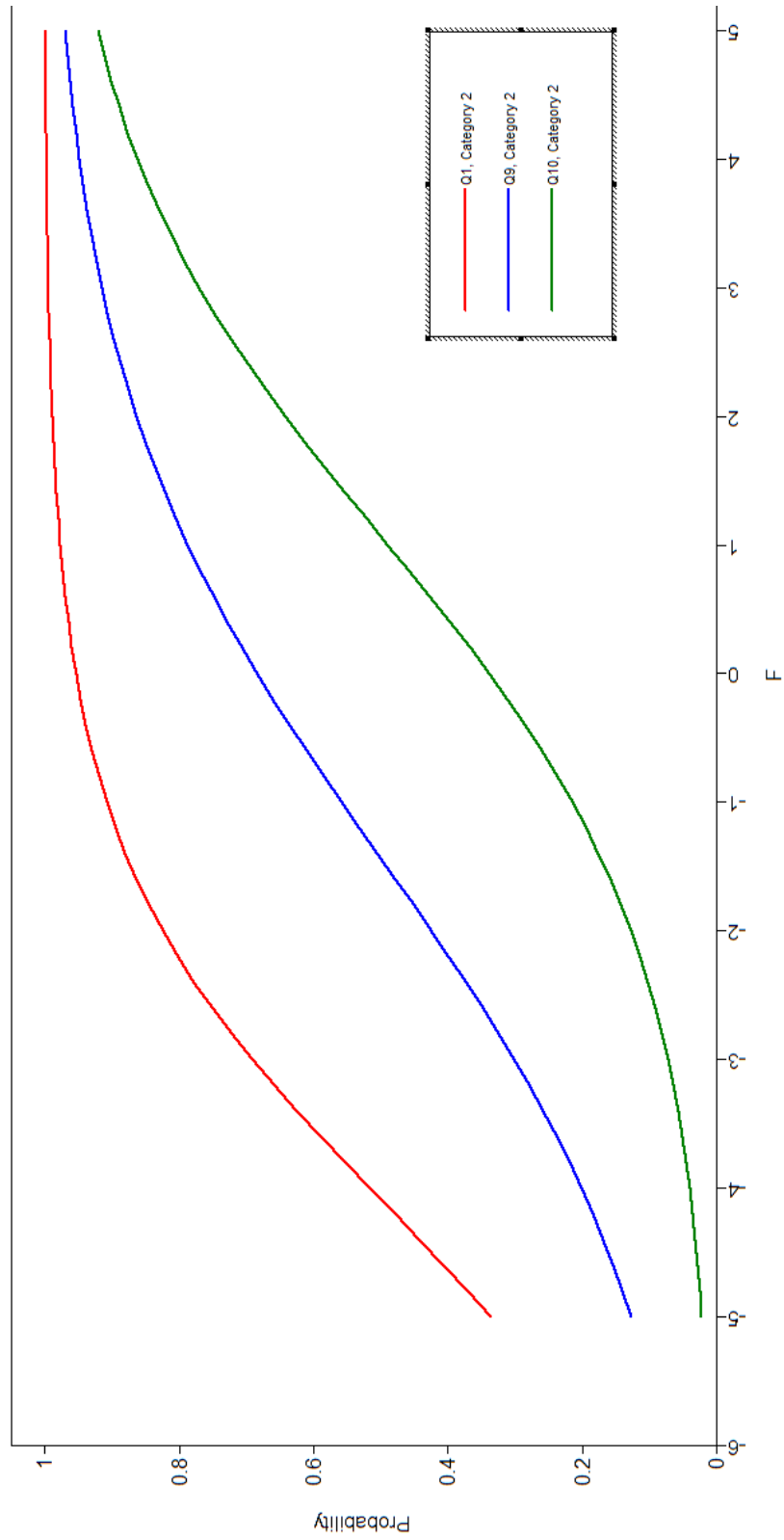


**Figure 12. One Latent Factor with 20 Predictors Model**

Our latent variable  $f$  represented by a circle is statistical knowledge. Latent variables are not able to be directly measured but through proper test construction we can attempt to estimate them, in this case we are wanting to find learners statistical knowledge. The 20 squares are the 20 items that are measurable such as the 20 questions used on either the original or transfer test within this project. After the IRT data analyses is computed a 2-parameter-logistical (or 2PL) model will result in

returning item difficulties and item discriminations. Item difficulties are measuring how challenging the respective questions are and item discriminations are essentially a validity measurement that tells us how well the respective item is at measuring what it is that we are attempting to measure.

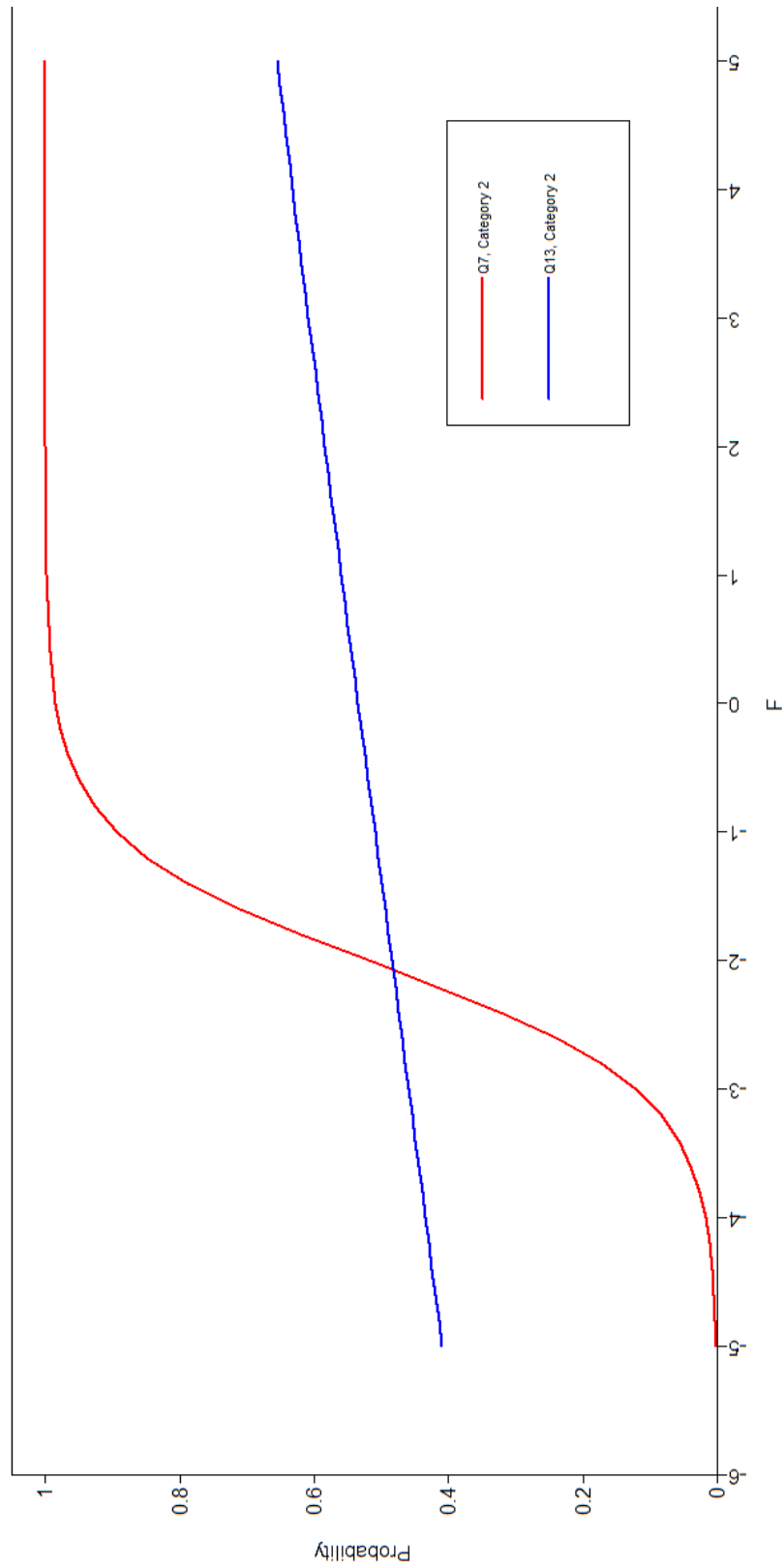
Though there are more complex IRT models available, the 2PL model is adequate in demonstrating the usefulness of Item Factor Analysis. The models that will be used below will also be using dichotomous response variables (e.g. I will only be analyzing the multiple choice data). When using IRT psychometricians often look at Item Characteristic Curves (or ICCs). ICCs give researchers a visualization of information about questions, each line represents a different question. Item difficulties will change by shifting left or right on the graph. In the following examples easier items will be shifted to the left while more difficult items will be shifted to the right. As can be seen in Figure 13, Q1 is an easy question while Q10 is a challenging question (the x-axis represents the latent trait of statistical knowledge, while the y-axis represents the probability of getting the answer correct).



**Figure 13. Two-Parameter-Logistical Item Response Theory Model with Varying Item Difficulties**



The ICC's represented in Figure 13 and Figure 14 are 2PL IRT models. The slopes of the ICC's depicted within these Figures represent item discriminations, typically as long as we have enough data present the steeper the slope the better that respective item will discriminate. In other words the steeper the slope the more valid the question is at representing the respective latent trait we are attempting to estimate. Again, in this project the latent trait we are attempting to measure is statistical knowledge. Items with flat or inverted slopes are either poor questions (e.g. have low validity) or are measuring something else (e.g. something other than statistical knowledge). See Figure 14.

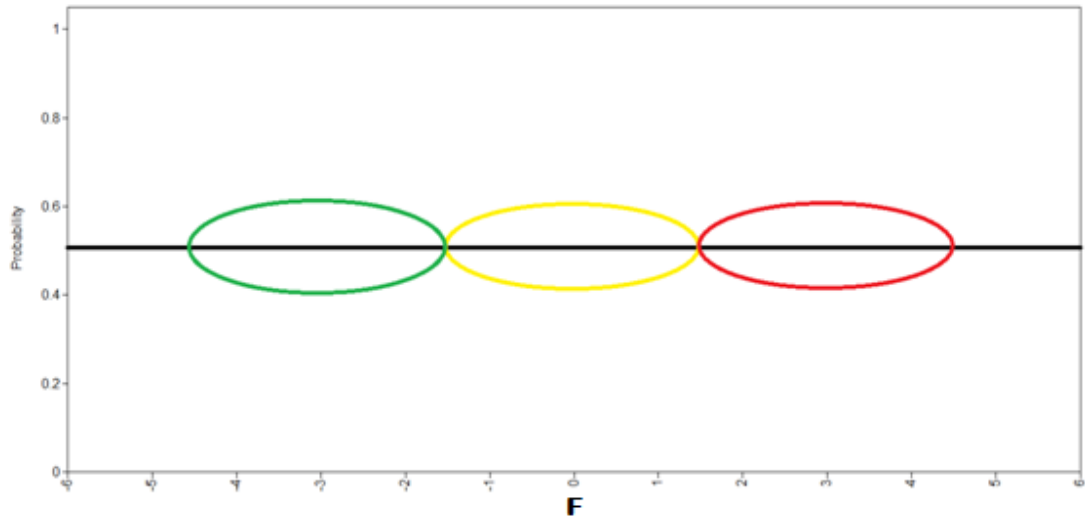


**Figure 14. Two-Parameter-Logistical Item Response Theory Model with Varying Item Discriminations**

Depicted in Figure 14 are the best and worst questions used within the original set of 20 questions within this study. Q7 has one of the steepest slopes resulting in good item discrimination (e.g. question with good validity at measuring statistical knowledge). While Q13 has flat slope resulting in poor item discrimination (e.g. question with low validity at measuring statistical knowledge). If learners answer Q7 correct they are likely to have a moderate or higher knowledge of statistical understanding (poor performers are getting this question wrong while moderate and high performers are getting this questions right). While if learners answer Q13 correctly it tells us very little about their understanding of statistics; poor, moderate, and high performers almost all have an equal chance of getting this question correct. In future examination it would be advised to keep question 7 but omit or replace question 13 for a different item with better discrimination / validity.

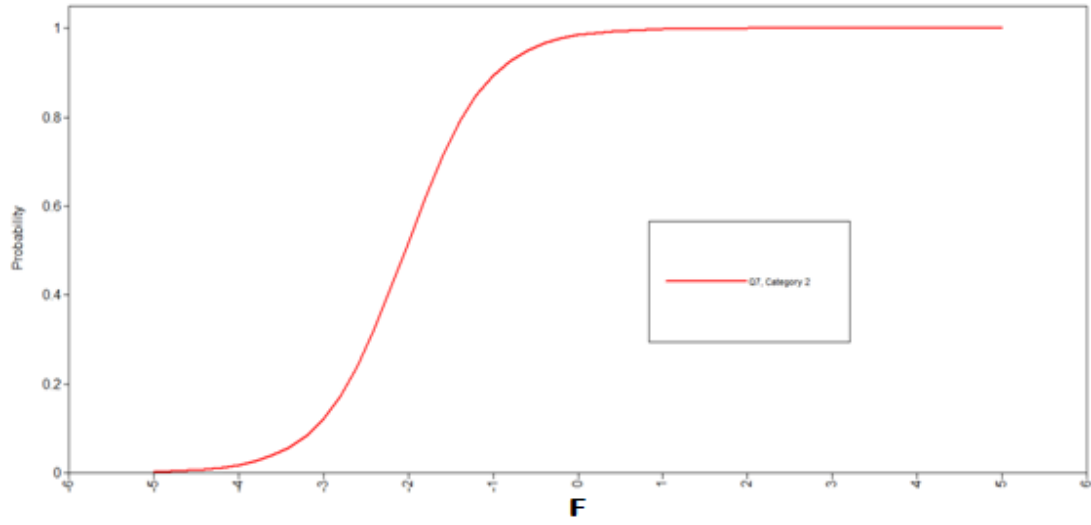
In theory what might a "good" test look like? What would the item difficulties and item discriminations yield? In terms of item difficulty we want to be able to measure individuals of all varying degrees of the respective latent trait in which we are attempting to measure. For example we would want an equal number of easy, moderate, and challenging questions in order to better estimate where each learner's knowledge level really is. Ideally we would want ICC's that intersect the .5 probability threshold at all possible latent trait levels in order to better estimate learners knowledge. Figure 15 depicts where we would want to see each of our items intersect the 50% threshold. Easy questions would be spread out in the green circle, moderate questions would be dispersed within the yellow, and challenging questions would be given within the red

circle. Low performers will likely only get green questions correct while high performers will get green, yellow, and some of the red questions correct.



**Figure 15. Hypothetical Item Difficulties for a "Good" Test**

How would item discriminations look on a “good” test? In a perfect world the items on a test would discriminate almost perfectly. That is each of a varying ICCs would have a steep slope that are almost identical. One could imagine one of the best questions used in this project depicted in Figure 16 populated across all possible latent levels (depicted in Figure 15). Such an exam with steep item discriminations across all varying latent difficulties would in theory result in a great tool for estimating one’s respective latent ability. The exam does not have to be lengthy in construction; rather it just needs to be made properly with quality controlled questions. IRT allows for such a test to be made if researchers and teachers are willing to put in the effort necessary to construct such a test.



**Figure 16. Hypothetical Item Discriminations for a "Good" Test**

## Chapter 5: Application of Item Response Theory

During this part of the project I wanted to highlight the usefulness of IRT within the current project and show how it can be used to improve test construction. It should be noted that IRT models are essentially factor analysis type models and thus require a large sample size to be utilized correctly (an  $N \geq 300$  or 400 is ideally preferred to assure that the IRT model is robust) (Reise & Revicki, 2015). The 2PL (2 parameter logistical) models depicted below both have relatively small sample sizes and thus interpretations of these models must be taken with caution. Each of the 2PL IRT models analyzed below are returning item difficulties and item discriminations (Embretson & Reise, 2000). Both models below are unidimensional and are taking in dichotomous dependent variables (correct vs incorrect). Each individual question may be thought of as a logistical regression with the respective item discrimination representing the respective coefficient associated with that question. As a whole the 2PL IRT model may also be thought of as a confirmatory factory analysis (or CFA). The respective questions that load highly, in theory will be more valid in measuring the respective latent trait we are attempting to measure (Embretson & Reise, 2000). Both models below were estimated using maximum likelihood within mPlus 7. The IRT analysis for the original 20 question used in the project within a multiple choice format can be seen in Table 14.

**Table 14. Two-Parameter-Logistical Model for Statistical Questions in Experiment 1**

( <i>N</i> = 245)	Correct	Percent Correct	Item Difficulty	Item Discrimination
Q1	231	94%	-4.09	0.74*
Q2	208	84%	-1.87	1.13***
Q3	224	91%	-2.00	1.65**
Q4	153	62%	-0.90	0.61**
Q5	214	87%	-2.18	1.06***
Q6	217	87%	-2.21	1.12**
Q7	230	94%	-2.04	2.05**
Q8	216	88%	-1.74	1.62***
Q9	165	67%	-1.43	0.54**
Q10	86	35%	1.07	0.63**
Q11	116	47%	0.13	0.99***
Q12	112	46%	0.44	0.40
Q13	131	54%	-1.38	0.10
Q14	223	91%	-3.22	0.79*
Q15	160	65%	-1.04	0.67**
Q16	134	55%	-1.21	0.16
Q17	232	95%	-2.15	2.01***
Q18	200	82%	-1.67	1.09***
Q19	224	92%	-1.98	1.67***
Q20	195	80%	-1.50	1.12***

\**sig p* = .05, \*\**sig p* ≤ .01, \*\*\**sig p* ≤ .001

The 2PL IRT model yields that overall the original 20 question test is somewhat easy with 17 out of 20 of the questions loading significantly on statistical knowledge. Item difficulties can be seen within the item difficulty column. Low negative numbers indicate easy items (e.g. questions 1, 6, 14) while numbers around 0 or greater indicate moderate or hard questions (e.g. questions 10, 11, 12). The three items (12, 13, 16) that loaded poorly are either bad questions (e.g. poor validity) or are measuring a different latent trait other than statistical knowledge (Reise & Henderson, 2003). It is advised in the future to replace these respective questions if this test is to be used again in the future. The ICCs for all 20 questions can be seen in Figure 17.

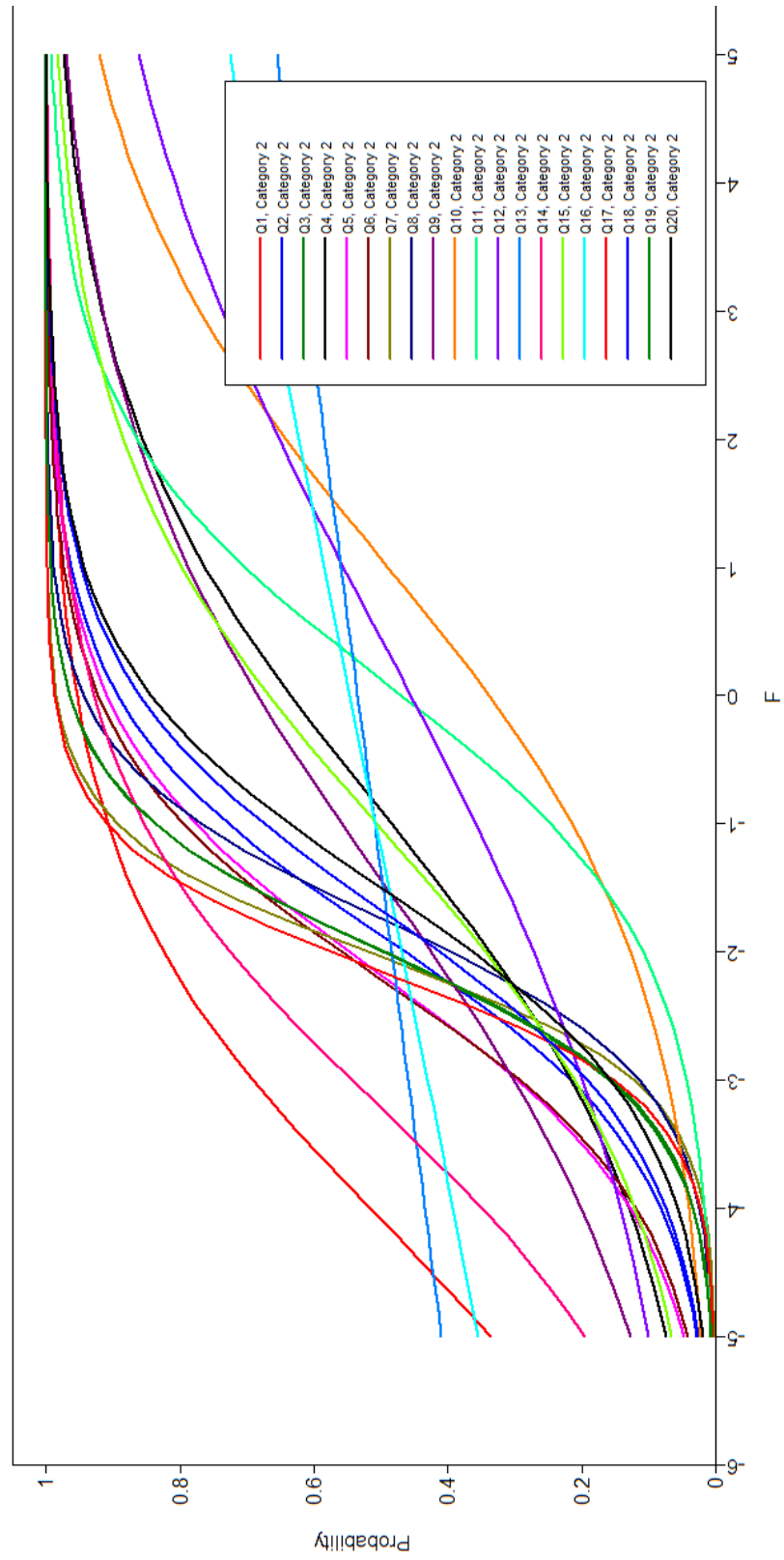


Figure 17. Item Characteristic Curves for the Original Statistical Test



In theory the x-axis in figure 17 is measuring statistical knowledge while the y-axis tells us the probability of an individual getting the answer correct given their statistical knowledge. Individual's with a latent level between -4 and -2 will likely do poorly on this exam while individuals with a trait level of 0 to 2 will likely do well on this exam. Within figure 17 easier questions are shifted to the left while harder questions are shift to the right. Also, notice that questions with higher loadings (e.g. questions 7 and 17) tend to have steep slopes while questions with low loadings tend to have flatter slopes (e.g. questions 12, 13, 16). The 2PL IRT model for transfer questions within a multiple choice format can be seen in Table 15.

**Table 15. Two-Parameter-Logistical Model for Transfer Questions in Experiments 2 & 3**

(N = 81)	Correct	Percent Correct	Item Difficulty	Item Discrimination
Q1	42	52%	-0.09	1.09**
Q2	69	85%	-1.41	1.94
Q3	72	89%	-1.82	1.58**
Q4	77	95%	-2.26	1.89
Q5	73	90%	-2.52	1.03
Q6	62	77%	-1.43	0.98*
Q7	80	99%	-2.20	7.68
Q8	64	79%	-1.48	1.10*
Q9	40	49%	0.05	0.49
Q10	56	69%	-1.43	0.61
Q11	18	22%	3.01	0.43
Q12	26	32%	1.53	0.52
Q13	57	70%	-0.95	1.15*
Q14	69	85%	-1.41	1.92*
Q15	68	84%	-1.49	1.55*
Q16	23	29%	1.35	0.77*
Q17	71	88%	-2.48	0.90*
Q18	69	85%	-1.73	1.30*
Q19	52	64%	-0.63	1.21**
Q20	59	72%	-0.84	1.86**

\*sig p = .05, \*\*sig p ≤ .01, \*\*\*sig p ≤ .001

Results of the 2PL IRT model depicted within Figure C must be interpreted with care due to the relatively small sample sized used. Compared to the original exam it can be seen that the transfer questions are typically more difficult (e.g. item difficulties in general are typically greater). Also, it can be seen that the transfer quiz in general seems to be less valid than the original quiz. Eight out of the 20 questions are not significantly loading onto our statistical knowledge factor suggesting that only 12 of our 20 questions are actually measuring what we are attempting to estimate within the transfer quiz. Typically it would be advised to replace the 8 questions with new questions with higher item discriminations. However, due to the low sample size used within this analysis it would likely be better to gather more data before reconstructing the test. The ICCs for all 20 transfer questions can be seen within figure 18.

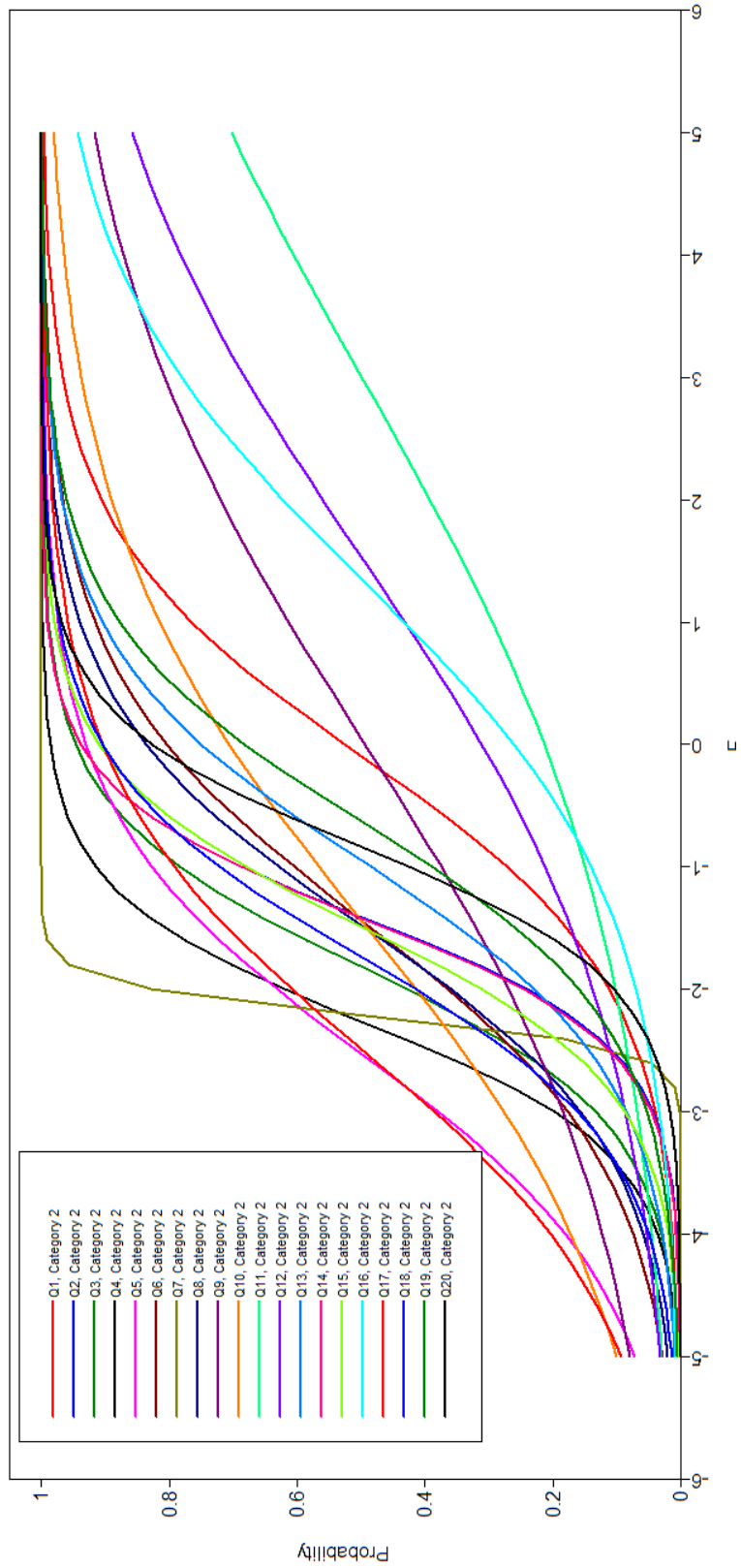


Figure 18. Item Characteristic Curves for the Transfer Test

Though the majority of the items in the transfer test do not discriminate as well as the items on the original test, the items on the transfer exam seem to have a greater dispersion of difficulty (the medians of the ICCs are spread out more). Arguably, greater median dispersion is quality to strive for in the attempt to construct a “good” test. Exams with greater median difficulty dispersions of the ICCs will allow exams to measure individuals differently that have different latent trait levels (Embretson & Reise, 2000); ideally we want tests that are able to do this (a good exam should be able to differentiate a fantastic performer from a mediocre performer). The ICCs of the transfer test also illustrate the larger number of items that discriminate poorly; these ICCs tend to have flatter slopes (e.g. questions 9 through 12). Again, all interpretations of the 2PL IRT model and its ICCs for the transfer test should be taken with caution due to the small sample size. If the sample size were to be quadrupled the item difficulties and item discrimination would be more representative of what they actually are in nature. As with most statistical tests larger samples sizes are typically favored due to greater statistical power, IRT models are no different (Reise & Revicki, 2015).

## Chapter 6: General Discussion

The main hypotheses investigated within this project as well as the evidence that was found to support or deny those respective claims can be found in Table 16.

**Table 16. Outcome of Project Hypotheses**

Hypotheses	Evidence
<b>Experiment 1</b>	
<ul style="list-style-type: none"> <li>• Retrieval practice conditions will outperform the instruction only condition.</li> </ul>	Supported
<ul style="list-style-type: none"> <li>• Open ended and multiple evaluation formats will behoove learning more than a multiple choice format.</li> </ul>	Not Supported
<ul style="list-style-type: none"> <li>• Conditions that are segmented will outperform conditions that are not segmented.</li> </ul>	Not Supported
<ul style="list-style-type: none"> <li>• An additive learning effect between the retrieval practice and segmentation procedures are expected.</li> </ul>	Not Supported
<ul style="list-style-type: none"> <li>• Multiple evaluation type formats will have greater insight into their own knowledge.</li> </ul>	Supported
<b>Experiment 2</b>	
<ul style="list-style-type: none"> <li>• Retrieval practice with feedback conditions will outperform retrieval practice and instruction only conditions.</li> </ul>	Supported
<ul style="list-style-type: none"> <li>• Conditions that are segmented with feedback will outperform conditions that are not segmented.</li> </ul>	Not Supported
<ul style="list-style-type: none"> <li>• An additive learning effect between the retrieval practice and segmentation procedures are expected.</li> </ul>	Not Supported
<ul style="list-style-type: none"> <li>• Retrieval practice with feedback conditions will outperform instruction only conditions in terms of transfer.</li> </ul>	Not Supported
<ul style="list-style-type: none"> <li>• The added manipulation of corrective feedback will boost all learners' insight into their own knowledge.</li> </ul>	Mixed Results
<b>Experiment 3</b>	
<ul style="list-style-type: none"> <li>• Retrieval practice conditions will outperform the instruction only condition in terms of transfer knowledge.</li> </ul>	Mixed Results
<ul style="list-style-type: none"> <li>• Conditions that are segmented will outperform conditions that are not segmented in terms of transfer.</li> </ul>	Not Supported
<ul style="list-style-type: none"> <li>• Multiple evaluation type formats will have greater insight into their own transfer knowledge.</li> </ul>	Not Supported

\*Evidence indicated within this table includes the general trend of statistical significance found in Experiments 1-3.

The robustness of the retrieval practice phenomenon can be seen within a multitude of experiments administered in a variety of fashions (Rowland, 2014). As such, the learning value of testing as a mechanism for learning has been seen throughout this project as well. Experiment 1 provides evidence that retrieval practice may be utilized using complex statistical material within an automated instructional environment. Providing learners with questions before a final examination tended to boost their performance by about a letter grade. Experiment 2 suggests that the retrieval practice phenomenon can be enhanced by utilizing immediate corrective feedback. Providing participants with immediate corrective feedback tended to behoove their performance another letter grade on top of retrieval practice alone. Thus, by utilizing both questions and immediate corrective feedback, student's knowledge acquisition may be enhanced by about 2 letter grades according to the data seen within this project.

Experiment 3 suggests that retrieval practice has limited success when transferring the learned knowledge into novel situations. Though one condition did seem to favor retrieval practice when administered the transfer quiz (the retrieval practice only condition within a multiple choice format), for the most part the benefits associated with retrieval practice did not seem to transfer. As with the results found here, the literature also suggests that benefits of retrieval practice may or may not be beneficial for knowledge transfer. Campbell and Robert (2008) as well as Zhou, Ma, Li, and Cui (2013) found both positive and null results when learners were given a knowledge transfer task after learning with retrieval practice. Though it may be somewhat more difficult to administer, the key may be to actually practice knowledge transfer with novice learners in order to improve their effectiveness of transferring their

known knowledge to novel situations. Another reason that we found no difference between the various learning conditions and knowledge transfer is that the transfer test could have been either too difficult or it was poorly constructed. Recall that the transfer quiz was perceived to be more difficult than the original quiz regardless of format. It was found, that 8 out of the 20 questions used on the transfer test were not valid; though the sample size was fairly small to properly construct the IRT models it could also be an indicator of poor test construction. Had the transfer test been better constructed, difference between the learning conditions may have been seen.

Unlike prior research segmentation had no impact on learning within this project (Mayer, 2001; Mayer, 2008; Mayer & Moreno, 2003). In a series of experiments Mayer (2008) reported medium to large effect sizes found within multimedia learning environments. The difference in these findings are likely due to differences within the procedure and differences of the stimuli used within the experiments. Mayer (2008) reported using complex science stimuli while this project used complex statistical stimuli. The stimuli differences as well as the procedural differences are likely why no segmentation effect was found. Arguably such a stipulation may point to a lack of robustness found with the segmentation phenomenon.

For the most part throughout the experiment participants within this study did a fairly good job at predicting their own performance. In experiment 1 it was found that conditions that utilized an open ended or multiple evaluation type format tended to have greater insight into their own knowledge. While learners within a multiple choice type format were able to significantly predict their own performance they did not do it as well as learners that were administered the information in open ended or multiple

evaluation type formats, as indicated by the respective effect sizes reported above. Open ended formats typically elicit a cued recall response which in turn is more difficult than its recognition memory counterparts (Rowland, 2014). It is thought that the added difficulty of a cued response aided learner's perception of their own knowledge. While the multiple evaluation format elicited a recognition memory response the learners were actively engaged in a judgment of learning task when administered the questions in this fashion. It is expected that metacognitive benefit associated with the multiple evaluation format is due to actually practicing a judgment of learning task. Engaging in a judgment of learning task is thought to aid learner's metacognitive ability even though the task involves recognition memory. A decrease in metacognitive performance was seen within the multiple choice condition because that respective condition utilized recognition memory without the use of a judgment of learning type task. In short, the multiple choice type format is easy which can lead learners to have a greater bias in what they think they know or do not know. Because the open ended and multiple evaluation type formats were thought to be more effortful it was originally thought that those formats would result in greater overall knowledge retention, however the benefits associated with the added effort seem to aid learners metacognitive insights instead.

In experiment 2 feedback was added to all conditions. The addition of immediate corrective feedback was beneficial to both learning and metacognitive insight. If learners failed at recalling the correct information at test they were immediately reminded of the correct answer in the form of feedback. This immediate corrective feedback, though quick, allows learners another study opportunity to relearn the information that they may not know so well. In addition, immediate corrective



feedback provides learners with direct insight into what they know or do not know by providing them with the correct information. Learners typically are then able to affirm a hit thereby boosting confidence or correct a miss thereby relearning the new information.

Item Response Theory is a powerful test building tool that can be used to make tests more efficient and more valid. The main models shown within this project illustrated how item difficulties and item discriminations can be taken into consideration when building a test. In theory a good test would have a wide array of questions with varying item difficulties in order to capture most individual's performance levels on the respective latent trait we are attempting to estimate. While item discrimination on a theoretically sound test would be relatively large with steep slopes indicating that our questions on the respective measurement tool is indeed measuring what it is that we are attempting to measure (Embretson & Reise, 2000). With a wide array of questions with varying item difficulties and questions with valid item discrimination better tests and measurement devices may be constructed. IRT is modeling process that provides teachers and psychometricians alike the ability to better construct valid tests and valid measurement devices. Utilization of IRT may be expensive on the front end of projects but at least we have evidence that our measurement devices are measuring what we indeed think they are measuring.

Though the findings from this project found no evidence of a retrieval practice + segmentation effect this research was still necessary to carry out. Perhaps instead of segmentation, learners and future research may want to investigate the impact of spaced learning in conjunction with retrieval practice. In terms of optimal learning, the spaced

testing effect has been found to be quite effective in other learning domains (Carpenter & DeLosh, 2005; Maddox & Balota, 2015). It is likely that the spaced testing effect would be beneficial for automated statistical learning as well.

The results from this project suggest that teachers as well as instructional designers should utilize testing and corrective feedback when possible. Testing is not just a method in which to measure individual's abilities but rather testing can be used as potent learning events (Karpicke & Roediger, 2007). Though multiple choice tests are better than no tests at all it is advised that open ended or multiple evaluation type questions be administered on examinations due to the metamemory benefits associated with those two formats. Metacognitive insight to such knowledge can lead learners in the correct direction as to where they should focus their restudy / retesting so that their time may be used appropriately. We as teachers and psychometricians should continue to strive to find new ways to teach and instruct students utilizing effective and efficient teaching methods and when it comes time to actually take that final test we owe it to ourselves and the learners to make sure that it is done properly with a test that has been well designed.

## References

- Baddeley, A. (1992). Working memory. *Science*, 255 (5044), 556-559.
- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(5), 1118-1133.
- Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, 13(4), 273-281.
- Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 918-928.
- Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, 36(3), 604-616.
- Campbell, J. I. D., & Robert, N. D. (2008). Bidirectional associations in multiplication memory: Conditions of negative and positive transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3), 546-555.
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19(5), 619-636.
- Dirkzwager, A. (1996). Testing with personal probabilities: 11-year-olds can correctly estimate their personal probabilities. *Educational and Psychological Measurement*, 56(6), 957-971.
- Dirkzwager, A. (2003). Multiple evaluation: A new testing paradigm that exorcizes guessing. *International Journal of Testing*, 3(4), 333-352.
- Dunlosky, J., Hertzog, C., Kennedy, M. R. F., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology*, 10(1), 4-11.
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, 21(4), 499-526.

- Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4), 801-812.
- Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review*, 17(6), 797-801.
- Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(1), 290-296.
- Huffman, W. B., & Hahn, S. (2015) Cognitive Principles in Mobile Learning Applications. *Psychology*, 6, 456-463.
- Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, 101(3), 621-629
- Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition*, 38(8), 1009-1017.
- Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57(2), 151-162.
- Kemp, N., & Grieve, R. (2014). Face-to-face or face-to-screen? undergraduates' opinions and test performance in classroom vs. online learning. *Frontiers in Psychology*, 5.
- Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, 138(4), 449-468.
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, 65(2), 85-97.
- Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomized controlled trial. *Medical Education*, 43(12), 1174-1181.
- Maddox, G. B., & Balota, D. A. (2015). Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty. *Memory & Cognition*, 43(5), 760-774.
- Maki, W. S., & Maki, R. H. (2002). Multimedia comprehension skill predicts differential outcomes of web-based and lecture courses. *Journal of Experimental Psychology: Applied*, 8(2), 85-98.

- Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.
- Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, *63*(8), 760-769.
- Mayer, R. E., Mathias, A., & Wetzell, K. (2002). Fostering understanding of multimedia messages through pre-training: Evidence for a two-stage theory of mental model construction. *Journal of Experimental Psychology: Applied*, *8*(3), 147-154.
- Mayer, R. E. & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*, 43–52.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437-447.
- Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*(2), 93-103.
- Reise, S. P., & Revicki, D. A. (Eds.). (2015). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York, NY: Routledge.
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432-1463.
- Skinner, B. F. (1961). Teaching machines. *Scientific American*, *205*, 90-112.
- Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 80-95.
- Zhou, A., Ma, X., Li, J., & Cui, D. (2013). The advantage effect of retrieval practice on memory retention and transfer: Based on explanation of cognitive load theory. *Acta Psychologica Sinica*, *45*(8), 849-859.

## Appendix A: The 20 Statistical Questions Used in Experiments 1 – 3

Questions	Possible Answers
<p><b>Q1:</b> There is a negative correlation between the number of pirates in the world and the amount of global pollution. A local researcher suggests that in order to decrease the amount of global pollution produced we should bring back pirates. Do you think this correlation meaningful?</p>	<p>a) Yes, the correlation is likely to be meaningful because as the number of pirates decrease pollution has increased.</p> <p>b) Yes, the correlation is likely to be meaningful because as the number of pirates increase pollution has also increased.</p> <p>c) No, though there may be a correlation in nature between pirates and the amount of global pollution, these two variables are likely not directly related.</p> <p>d) No, the correlation that actually exists in nature is likely to be positive, not negative.</p>
<p><b>Q2:</b> Is it possible to have a correlation of <math>r = -1.3</math>?</p>	<p>a) Yes, <math>r = -1.3</math> is within the possible range for Pearson's <math>r</math>.</p> <p>b) No, <math>r = -1.3</math> is not within the possible range for Pearson's <math>r</math>.</p> <p>c) Yes, <math>r = -1.3</math> is a strong positive correlation.</p> <p>d) No, Pearson's <math>r</math> cannot handle negative values.</p>
<p><b>Q3:</b> Which of the following correlations is stronger: <math>-0.4</math> or <math>0.4</math>?</p>	<p>a) <math>0.4</math> is a stronger correlation than <math>-0.4</math>.</p> <p>b) <math>-0.4</math> is a stronger correlation than <math>0.4</math>.</p> <p>c) Both correlations are equal in strength.</p> <p>d) Not applicable. Pearson's <math>r</math> cannot handle negative values.</p>
<p><b>Q4:</b> Some researchers at a local university have found that there is a correlation between parental reading levels and their child's academic performance <math>r = 0.45</math>. What is the direction of this correlation and strength of this correlation?</p>	<p>a) Strong correlation in the negative direction.</p> <p>b) Weak correlation in the negative direction.</p> <p>c) Medium correlation in the positive direction.</p> <p>d) Strong correlation in the positive direction.</p>
<p><b>Q5:</b> Some researchers at an international university have found that there is a</p>	<p>a) Strong negative correlation.</p> <p>b) Weak negative correlation.</p>

correlation between age of a child and their mathematical ability, $r = -.09$ . What is the direction of this correlation and strength of this correlation?	<ul style="list-style-type: none"> <li>c) Weak positive correlation.</li> <li>d) Strong positive correlation.</li> </ul>
<b>Q6:</b> We are curious if a new supplement is helping males gain physical strength. Physical strength is measured by the amount individuals can lift during certain strength building exercises. Researchers are testing the new supplement compared to a placebo. What are the independent and dependent variables in this experiment?	<ul style="list-style-type: none"> <li>a) The independent variable is the amount an individual can lift during the strength test; The dependent variable is supplement type (real vs. placebo).</li> <li>b) The independent variable is supplement type (real vs. placebo); The dependent variable is the amount an individual can lift during the strength test.</li> <li>c) The independent variable is supplement type (real vs. placebo); The dependent variable is the participant's heart rate.</li> <li>d) Not enough information is provided to answer the question.</li> </ul>
<b>Q7:</b> How many factors does the following design have: $2 \times 2 \times 3 \times 3$ ?	<ul style="list-style-type: none"> <li>a) 2</li> <li>b) 3</li> <li>c) 4</li> <li>d) 36</li> </ul>
<b>Q8:</b> How many levels are within the fourth factor in the following experimental design: $2 \times 4 \times 8 \times 16$ ?	<ul style="list-style-type: none"> <li>a) 2</li> <li>b) 4</li> <li>c) 8</li> <li>d) 16</li> </ul>
<b>Q9:</b> If we need 10 participants for each condition (level), how many participants do we need for the following design: $2$ (between) $\times 2$ (between)?	<ul style="list-style-type: none"> <li>a) 10</li> <li>b) 20</li> <li>c) 40</li> <li>d) Answer not provided.</li> </ul>
<b>Q10:</b> If we need 40 participants for each condition (level), how many participants do we need for the following design: $2$ (within) $\times 2$ (within)?	<ul style="list-style-type: none"> <li>a) 40</li> <li>b) 80</li> <li>c) 120</li> <li>d) 160</li> </ul>
<b>Q11:</b> We have a barometer (an instrument that measures atmospheric pressure) that consistently gives us the wrong measurement. Is this device reliable? Is this device valid?	<ul style="list-style-type: none"> <li>a) The device is both reliable and valid.</li> <li>b) The device is reliable but not valid.</li> <li>c) The device is valid but not reliable.</li> <li>d) The device is neither reliable nor valid.</li> </ul>
<b>Q12:</b> We conduct an experiment within the lab and do a thorough job of controlling all variables included within the experiment. Which types of validity are likely to be high?	<ul style="list-style-type: none"> <li>a) Internal validity will likely be high but external validity will likely be low.</li> <li>b) Internal validity will likely be low but external validity will likely be high.</li> <li>c) Both internal and external validity are likely to be high.</li> </ul>

	d) Both internal and external validity are likely to be low.
<b>Q13:</b> We conduct a field study and do a very poor job of controlling for all variables used in our experiment. Which types of validity are likely to be high?	<p>a) Internal validity will likely be high but external validity will likely be low.</p> <p>b) Internal validity will likely be low but external validity will likely be high.</p> <p>c) Both internal and external validity are likely to be high.</p> <p>d) Both internal and external validity are likely to be low.</p>
<b>Q14:</b> We are interested in measuring three different schools ACT scores. The school with the highest mean score on the ACT gets a free pizza day towards the end of the semester. Should the three different schools be treated as a categorical or continuous variable? Should the ACT scores be treated as a categorical or continuous variable?	<p>a) Both school and ACT score should be treated as a continuous variable.</p> <p>b) Both school and ACT score should be treated as a categorical variable.</p> <p>c) Schools should be treated as a continuous variable and ACT score should be treated as a categorical variable.</p> <p>d) Schools should be treated as a categorical variable and ACT score should be treated as a continuous variable.</p>
<b>Q15:</b> In modern day professional sports it is often thought to be the case that the taller (inches) and faster (40 meter dash) someone is the better they will perform in sports. An experiment is designed to further investigate this question. Should size be treated as a categorical variable or continuous variable? Should speed be treated as a categorical variable or continuous variable?	<p>a) Both size and speed should be treated as continuous variables.</p> <p>b) Both size and speed should be treated as categorical variables.</p> <p>c) Size should be treated as a categorical variable and speed should be treated as a continuous variable.</p> <p>d) Size should be treated as a continuous variable and speed should be treated as a categorical variable.</p>
<b>Q16:</b> In 2014 the estimated life expectancy is the following for the countries indicated: Japan 84 years of age, United States 80 years of age, and Nigeria 53 years of age. Assuming that the respective governments made little or no errors when conducting this research on every death within their countries, what do these numbers represent?	<p>a) The numbers for each countries life expectancy represents <math>\bar{x}</math>.</p> <p>b) The numbers for each countries life expectancy represents <math>\mu</math>.</p> <p>c) The numbers for each countries life expectancy represents SD.</p> <p>d) The numbers for each countries life expectancy represents <math>\sigma</math>.</p>
<b>Q17:</b> The following section was taken from an experimental paper:  For the remaining 80 participants 64 were female and 16 were male with ages ranging from 18 – 22 ( $\bar{x}$ = 18.78,	<p>a) Sample mean = 18; Sample standard deviation = 22</p> <p>b) Sample mean = 64; Sample standard deviation = 16</p> <p>c) Sample mean = 18.78; Sample standard deviation = 0.98</p>

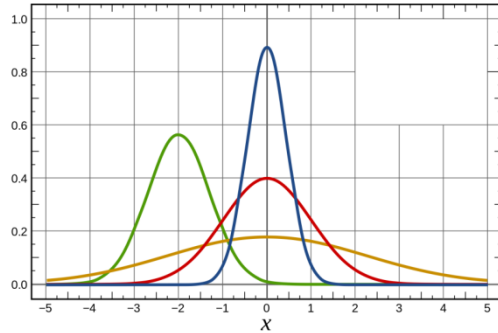


SD = 0.98).

What is the indicated sample mean and sample standard deviation?

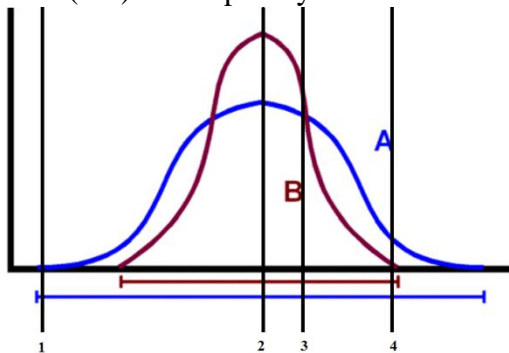
d) The data asked for is not listed in the section provided.

**Q18:** Which distribution (represented by different colors) has the lowest amount of variability?



- a) The green distribution.
- b) The yellow distribution.
- c) The red distribution.
- d) The blue distribution.

**Q19:** In theory if random sampling is done properly, where should the population mean fall within these distributions? Indicate a line number (1-4) and explain your answer.



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

**Q20:** Say we construct a 90% confidence interval for our sample mean. What is the likelihood that the confidence interval contains the population mean?

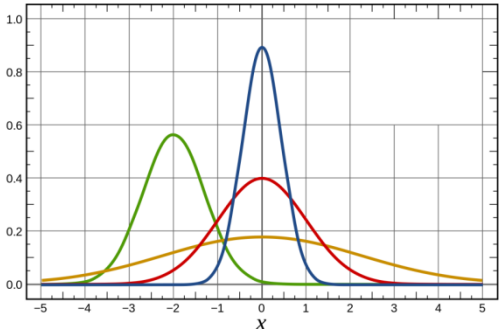
- a) 90%
- b) 10%
- c) 80%
- d) 95%

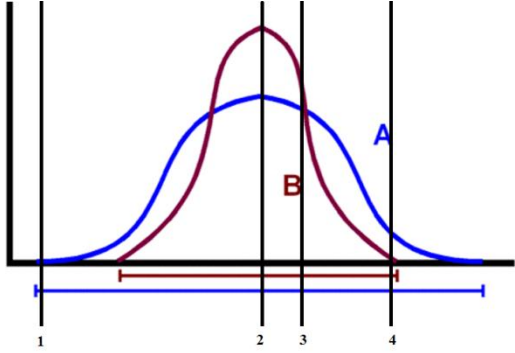
## Appendix B: The 20 Transfer Questions Used in Experiments 2 and 3

Questions	Possible Answers
<p><b>TQ1:</b> Some researchers that work for NSF (the National Science Foundation) found that there is a strong positive correlation between the amount of time spent studying for an exam and exam performance. The researchers conclude that if students wish to receive higher grades on examinations they should then spend more time studying for that respective exam. Do you think this correlation is meaningful?</p>	<p>a) Yes, this correlation is meaningful. It is also the case that time spent studying for an exam is causing higher exam performance.</p> <p>b) Yes, this correlation is meaningful. However, because this is a correlation no causal inferences may be made about the variables relationship at this time.</p> <p>c) No, this correlation is not meaningful. The two indicated variables are likely to have a strong negative relationship.</p> <p>d) No, this correlation is not meaningful. A third variable is likely impacting both variables ultimately causing the two measured variables to have a spurious correlation in nature.</p>
<p><b>TQ2:</b> Is it possible to have a correlation of <math>r = 0.0</math>?</p>	<p>a) Yes, <math>r = 0.0</math> is an indicator of a strong relationship between variables.</p> <p>b) Yes, <math>r = 0.0</math> is an indicator of a weak relationship between variables.</p> <p>c) Yes, <math>r = 0.0</math> is an indicator of a no relationship between variables.</p> <p>d) No, <math>r = 0.0</math> is reported only when something has gone wrong during our data analysis.</p>
<p><b>TQ3:</b> Which of the following correlations is stronger: <math>r = -0.8</math> or <math>r = 0.5</math>?</p>	<p>a) <math>r = -0.8</math> is stronger.</p> <p>b) <math>r = 0.5</math> is stronger.</p> <p>c) Both correlations are equal in strength.</p> <p>d) Not applicable. Pearson's <math>r</math> cannot handle negative values.</p>
<p><b>TQ4:</b> Researchers at a local community college found that there is a correlation between arousal levels and teacher evaluations, <math>r = 0.12</math>. What is the direction of this correlation and its strength?</p>	<p>a) Weak correlation in the negative direction.</p> <p>b) Strong correlation in the positive direction.</p> <p>c) Medium correlation in the negative direction.</p> <p>d) Weak correlation in the positive direction.</p>
<p><b>TQ5:</b> Researchers at Harvard found a strong positive relationship between socio-economic</p>	<p>a) <math>r = 0.32</math></p> <p>b) <math>r = -0.16</math></p>

<p>status (measure of family income) and a child's reading and math achievement on standardized tests. What is a possible Pearson's <math>r</math> value for this type of relationship?</p>	<p>c) <math>r = 0.64</math> d) <math>r = -0.50</math></p>
<p><b>TQ6:</b> An environment friendly energy company is curious which clean energy source is the most efficient. They are testing wind, hydroelectric, and solar all measured in terawatts (a unit of energy measurement). What are the Independent and Dependent Variables in this experiment?</p>	<p>a) The independent variable is the amount of force generated; The dependent variable is energy source type (wind, hydroelectric, and solar). b) The independent variable is the amount of terawatts generated; The dependent variable is energy source type (wind, hydroelectric, and solar). c) The independent variable is energy source type (clean and fossil); The dependent variable is amount of terawatts generated. d) The independent variable is energy source type (wind, hydroelectric, and solar); The dependent variable is amount of terawatts generated.</p>
<p><b>TQ7:</b> How many factors does the following design have: <math>4 \times 4</math>?</p>	<p>a) 2 b) 3 c) 4 d) 16</p>
<p><b>TQ8:</b> An I/O psychologist at OU was interested in the likelihood of bosses hiring various types of students as workers. The experimenter manipulated Dress (casual, business casual, or business), Posture (perfect, okay, slouching, or unconscious), and use of Filler Words (yes or no), as well as verifying that there was no prejudice against males or females (Gender). How many levels are within the third factor?</p>	<p>a) 2 b) 3 c) 4 d) 48</p>
<p><b>TQ9:</b> If we need 30 participants for each condition (level), how many participants do we need for the following design: <math>2</math> (within) <math>\times 3</math> (between)?</p>	<p>a) 30 b) 60 c) 90 d) 180</p>
<p><b>TQ10:</b> You are being paid by the University to tutor individuals that are struggling in statistics classes. A student comes to you and they are struggling with the differences that exist in between and within subject's designs.</p>	<p>a) In between subject's design each participant is only in one condition; while within subject's design each participant is in all conditions. b) Back in the day there was once a</p>

<p>How do you explain between and within subject designs to this learner?</p>	<p>difference between the terminology, however there is no longer a difference between the two terms.  c) The most important thing to remember is that both between subjects and within subject designs require the same amount of people in order to conduct our experiment properly.  d) In between subject's design all participants are in all conditions; while within subject's design each participant is in only one condition.</p>
<p><b>TQ11:</b> During the space race against the USSR, the United States sent up a scale in the shuttle with the astronauts to see what their weight would be while in space. However, due to an accident the scale was potentially damaged in the launch sequence. On earth the scale was both reliable and valid, however now they astronauts cannot get a measurement to read on the scale at all while in space. Each of 4 astronauts tries 5 times each but cannot seem to get a read out on the scale. Is this device reliable? Is this device valid?</p>	<p>a) The device is both valid and reliable.  b) The device is reliable but not valid.  c) The device is valid but not reliable.  d) Measurement error, we have no data to answer this question.</p>
<p><b>TQ12:</b> We conduct an experiment in the field and do a good job controlling all variables within our power. Though some things don't go as planned, the majority of our observations are kept consistent. Which types of validity are likely to be high?</p>	<p>a) Both internal and external validity will be high.  b) Internal validity will be high but external validity will be low.  c) Internal validity will be low but external validity will be high.  d) Neither internal nor external validity will be high</p>
<p><b>TQ13:</b> Fill in the blanks.  Generally, _____ experiments tend to have high _____ and have large amount of control while _____ experiments tend to have high _____ but lack control.</p>	<p>a) Field : external validity : lab : internal validity  b) Field : internal validity : lab : external validity  c) Lab : external validity : field : internal validity  d) Lab : internal validity : field : external validity</p>
<p><b>TQ14:</b> Dr. Kreiger is convinced that individuals with blood type O are healthier than individuals with other blood types. To test this he is comparing all blood types (A,</p>	<p>a) Blood type should be categorical and amount of glucose metabolized continuous.  b) Amount of glucose metabolized</p>

<p>B, AB, and O) in terms of how fast they metabolize glucose (measured in milligrams per deciliter mg/dl). Should our blood type variable be categorical or continuous? Should the amount of glucose metabolized be considered categorical or continuous?</p>	<p>should be treated as categorical and blood type should be treated continuous.  c) Both variables should be treated as categorical.  d) Both variables should be treated as continuous.</p>
<p><b>TQ15:</b> In your introduction to psychology class we are interested in conducting an academic battle of the sexes. This research will compare Gender (Male or Female), Class Type (Freshmen, Sophomore, Junior, or Senior) and overall class Grade (0 to 100%). In terms of measurement, how should each of these factors be treated?</p>	<p>a) All factors should be treated as categorical.  b) Gender should be categorical, Class type should be continuous, and Grade should be continuous.  c) Gender should be categorical, Class type should be categorical, and Grade should be continuous.  d) Gender should be continuous, Class type should be continuous, and Grade should be categorical.</p>
<p><b>TQ16:</b> In 2013, researchers attempted to measure life expectancy using the first 500 deaths in a given state. This is the data that was collected: Hawaii (81.3 years), California (80.8 years), Illinois (79.0 years), and Alabama (75.4 years). What do these numbers represent?</p>	<p>a) The numbers for each states life expectancy represents <math>\bar{x}</math>.  b) The numbers for each states life expectancy represents <math>\mu</math>.  c) The numbers for each states life expectancy represents SD.  d) The numbers for each states life expectancy represents <math>\sigma</math>.</p>
<p><b>TQ17:</b> Which distribution (represented by different colors) has the most amount of variability?</p> 	<p>a) The green distribution.  b) The yellow distribution.  c) The red distribution.  d) The blue distribution.</p>
<p><b>TQ18:</b> What is a 95% confidence interval exactly?</p>	<p>a) A range of data in which we are 95% certain that the population mean falls within.  b) The likelihood that our findings are found due to chance.  c) The amount of error that is associated with our sampled mean.</p>

	<p>d) A range of data in which we are 5% certain that the mean sampled does not exist.</p>
<p><b>TQ19:</b> Say we conducted a decent experiment and gathered a large amount of data for two overlapping data sets. Which line is the least likely to represent the population mean?</p> 	<p>a) 1 b) 2 c) 3 d) 4</p>
<p><b>TQ20:</b> Why do scientists often use sampling methods rather than directly measure population parameters?</p>	<p>a) Scientists tend to only take population parameters seriously; sampling practices are taken as a joke. b) Often times measuring population parameters is too expensive (or difficult) when sampling methods are available. c) Only samples are used, the population parameters are never truly known. d) Often times the population parameters are available, very rarely do we actually have to do sampling.</p>