2018-06-01

Addressing Pre-Service Teachers' Misconceptions About Confidence Intervals

Kiya Lynn Eliason
Brigham Young University

Addressing Pre-Service Teachers' Misconceptions About Confidence Intervals

Kiya Lynn Eliason

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Steven R. Jones, Chair
Daniel K. Siebert
Douglas L. Corey

Department of Mathematics Education

Brigham Young University

ABSTRACT

Addressing Pre-Service Teachers' Misconceptions About Confidence Intervals

Kiya Lynn Eliason
Department of Mathematics Education, BYU
Master of Arts

Increased attention to statistical concepts has been a prevalent trend in revised mathematics curricula across grade levels. However, the preparation of secondary school mathematics educators has not received similar attention, and learning opportunities provided to these educators is oftentimes insufficient for teaching statistics well. The purpose of this study is to analyze pre-service teachers' conceptions about confidence intervals. This research inquired about statistical reasoning from the perspective of students majoring in mathematics education enrolled in an undergraduate statistics education course who have previously completed an introductory course in statistics. We found common misconceptions among pre-service teachers participating in this study. An unanticipated finding is that all the pre-service teachers believed that the construction of a confidence interval relies on a sampling distribution that does not contain every possible sample. Instead, they believed it is necessary to take multiple samples and build a distribution of their means. I called this distribution the Multi-Sample Distribution (MSD).

Keywords: statistical reasoning, statistics education, teacher preparation, confidence interval estimation, multi-sample distribution

ACKNOWLEDGEMENTS

I am sincerely humbled to finish graduate school. The faculty and friends I had in my program directly impacted my efforts to complete this thesis in significant ways. I am honored to have associated with many of the brightest minds in my field. Specifically, I would like to thank Dr. Steven Jones, Dr. Dan Siebert, and Dr. Doug Corey for providing such insightful feedback about my study and manuscript. I appreciate their encouragement and enthusiasm. A critical cheerleader for my study has been my dear friend and officemate, Konda Luckau. She was there to talk me out of quitting every time I wanted to give up.

This accomplishment was only possible because of the continual support of my loved ones. I am grateful to my parents, Marty and Carrie Heaton, for the love they have given to me and for the sacrifices they have made for me throughout my life. I am grateful to my siblings, Siera, Johanna, Joshua, and Izak for believing in me and for making me smile. I appreciate my grandparents, Steve and Judy Heaton and Craig and Mary Kessler, for the examples they set for me of how to use one’s education to serve others. Finally, I wish to express gratitude for my three favorite people: my daughter, Makaylee; my husband, Jake; and my God. They give me reasons to believe in a glorious future.

TABLE OF CONTENTS

TITLE PAGE i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS iv

LIST OF FIGURES vii

LIST OF TABLES viii

CHAPTER 1: INTRODUCTION AND RATIONALE 1

CHAPTER 2: LITERATURE REVIEW AND THEORETICAL FRAMEWORK 6

Literature Review 6

Theoretical Framework 15

Knowledge in Pieces 15

Defining Misconception 17

Student Noticing 21

Research Questions 21

CHAPTER 3: METHODOLOGY 23

Setting and Participants 23

Data Collection 25

Data Analysis 28

Coding Pretests 28

Coding Interviews 31

Identifying Readouts and Causal Net Elements 34

CHAPTER 4: RESULTS 37

Participant Pretest Data.. 37

Interview Themes 39

Misconception 1: A Confidence Interval Allows Us To Make Inferences About More Than the Mean of the Population. 41

1a. Evidence. 42

1b. Connections to Classroom Experience. 44

1c. Resolution. 47

Misconception 2: The Range Spanned By a Confidence Interval with a 95% Confidence Level is the Range that Contains 95% of the Population Data. 50

2a. Evidence. 50

2b. Connections to Classroom Experience. 53

2c. Resolutions. 56

Misconception 3: Calculations for Confidence Intervals are Not Based on Actual Sampling Distributions. 58

3a. Evidence. 60

3b. Connections to Classroom Experience. 67

3c. Resolutions 69

Misconception 4: Statisticians Must Take Multiple Samples in Order to Compute a Reliable Confidence Interval. 71

4a. Evidence. 72

4b. Connections to Classroom Experience. 73

4c. Resolutions 77

Expert Conception 1: Variability Exists in the Values of Sample Statistics. 81

CHAPTER 5: DISCUSSION... 83

Not Encapsulating the ASD Was A Cause of Confidence Interval Misconceptions 84

Suggestions for Future Research 85

Limitations and Directions for Future Research... 86

Conclusion 88

REFERENCES 89

APPENDIX A: PRETEST QUESTIONS.. 97

APPENDIX B: INTERVIEW PROTOCOL.. 99

LIST OF FIGURES

Figure 1. Anna's drawing of two overlapping normal distributions. 45

Figure 2. Anna's skewed population image. 48

Figure 3. Tiana's pretest responses. 52

Figure 4. Tiana's drawing of results between 70 and 85... 53

Figure 5. Tiana's drawing of a 90% confidence interval. 54

Figure 6. Sampling distribution images used in textbooks.. 55

Figure 7. Comparisons of the width of various ranges... 56

Figure 8. Danielle's drawing of overlapping distributions... 78

LIST OF TABLES

Table 1. *Misconceptions of confidence intervals* 9

Table 2. *Participant demographics* 25

Table 3. *Participant interview dates* 28

Table 4. *Levels of proficiency for coding student responses* 29

Table 5. *Pre-existing codes* 32

Table 6. *Misconceptions observed in interviews with participants* 40

Table 7. *Misconception 1 observations* 42

Table 8. *Comparisons to the population mean given conditions of normality or C.L.T.* 46

Table 9. *Misconception 2 observations* 50

Table 10. *Misconception 3 observations* 60

Table 11. *Likely causal nets demonstrated by Corinne* 62

Table 12. *Likely causal nets demonstrated by Tiana* 63

Table 13. *Likely causal nets demonstrated by Danielle* 64

Table 14. *Likely causal nets demonstrated by Ethan* 66

Table 15. *Parallel causal nets for σ and ASD* 69

Table 16. *Misconception 4 observations* 72

CHAPTER 1: INTRODUCTION AND RATIONALE

As the math department chair at a public high school during the initial years of the CCSS implementation, I was concerned that our department was not prepared to teach all the standards relating to statistics. Of the ten teachers in my department, eight voluntarily admitted to me that they did not feel as though they were competent with statistical concepts. Only one fellow teacher felt confident teaching statistical concepts beyond calculating the mean, median, and mode of a data set. Unfortunately, these circumstances were not unique to my school. According to the Conference Board of the Mathematical Sciences, most new high school teachers will require further coursework to be prepared to teach more than basic statistics (CBMS, 2001).

The intent of my study is to investigate statistical conceptions of pre-service math teachers. I have been especially interested in confidence intervals because confidence intervals are a popular means of reporting results (Harlow, 1997) and a student's understanding of confidence intervals can influence how they interpret the conclusions. How a student interprets a confidence interval can have ramifications in many facets of life (Franklin et al., 2015). Examples include the selection of a personal health insurance plan given that an individual would like to minimize their financial loss or the selection of a date for an outdoor event given that the host would like to minimize the risk of there being rain on that day.

Statistics education is important because of the data-driven nature of our world. We have the internet at our fingertips, and today's generation is being raised with an information overload unlike anything we have ever known. Since we live in a data-driven world, we are expected to make decisions based on data in our everyday lives. A cognitive ability to process statistics is critical for adults to successfully navigate a myriad of choices in life whether they be following media coverage of current events, making financial decisions, or assessing health risks (Franklin et al., 2015). Therefore, statistical literacy is necessary for intelligent citizenship for all people - not just those conducting formal research.

Because of the need to educate students about statistics, the trend manifest in State Standards is to place increasingly heavy emphasis on statistics and probability, particularly in the secondary grades (e.g. American Mathematical Society [AMS], 2001; Common Core State Standards Initiative [CCSSI], 2010; Mathematical Association of America [MAA], 2004; National Council of Teachers of Mathematics [NCTM], 2000). For example, in 1989 the NCTM published standards in grades 9 through 12 with six statistical concepts. In 2017, the NCTM standards for grades 9 through 12 expanded to include 19 statistical concepts. Examples of new standards include understanding how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference as well as identifying trends in bivariate data and finding functions that model the data or transform the data so that they can be modeled.

A teacher's pedagogical knowledge must be grounded in content to be useful in teaching (Shulman, 1986), so student achievement will depend on educators who are knowledgeable about the statistics content that is being added to curricula. Yet, this is a problem because pre-service teachers may not be receiving the training they need to develop understanding of statistics content. For example, many secondary level mathematics teachers may major in mathematics, and in many universities the courses in data analysis do not count toward a mathematics degree (Franklin et al., 2015). Learning opportunities provided to pre-service teachers are oftentimes insufficient for teaching statistics well (Franklin et al., 2015). Thus, many math teachers are not prepared to teach statistics which in turn can lead to many of our students being unprepared to reason about statistics as adults.

Because of this problem, researchers have recently begun to turn their attention to in-service and pre-service teachers' understanding of statistical content (Batanero, Burrill, & Reading, 2011). Some of this research has focused on documenting misconceptions teachers or students might have about foundational statistical concepts, like p-values (Goodman, 2008) and confidence intervals (Fidler, 2006; Kalinowski, 2010; Cumming & Finch, 2005). I want to further this work by investigating how certain misconceptions might be resolved for pre-service student teachers.

My preliminary research question was directed at conceptions of p-values as a measure of variation because of a dearth of information I perceived in literature pertaining to this topic. However, I decided against investigating p-values in favor of using confidence intervals because of the debate about whether p-values should continue to be used (e.g. Wasserstein & Lazar, 2016; Murtaugh, 2014; Hubbard & Lindsay, 2008). Confidence intervals are the most commonly recommended alternative to using p-values in tests of significance (Harlow, 1997). Medical journals reporting statistics reflect this consensus, with approximately 85% of ten leading medical journals in 2003 using confidence intervals to report their findings (Coulson, Fidler, & Cumming, 2005). Due to the cases that are outlined against p-values and the popularity of confidence intervals, I determined that research about confidence interval estimation was likely to be more valuable to the field. Thus, I focused on what is known about student conceptions related to confidence intervals as a measure of variability. Confidence intervals have been strongly advocated as the best strategy for reporting statistical results (APA, 2001). A wider use of confidence intervals has been recommended as a means of improving research communication due to the damaging over-reliance on the extensively misunderstood hypotheses tests (Cumming, Williams, & Fidler, 2004; Harlow, 1997). Confidence intervals appear in everyday situations.

Examples include Gallup poll reports such as that Americans' average daily spending is $104 ± $5 (McCarthy, 2017), that 46% ± 3.9% of Ukrainians are struggling to afford food (Lyons, 2017), and that 77% ± 4% of Americans believe that the country's moral values are declining (Norman, 2017). Healthy and productive citizenship, employment, and family life require adults to be able to use such data intelligently (Franklin et al., 2007). However, confidence intervals are not always properly interpreted and are prone to misconceptions (Fidler, 2006; Fidler, Thomason, Cumming, Finch, & Leeman, 2004). Hence, all students should learn to understand the ideas behind confidence intervals, but their learning opportunities are largely influenced by the proficiency of their teachers. As a result, it is worthwhile to investigate pre-service teachers' conceptions about confidence intervals.

My personal interest in this topic stems largely from the fact that I worked as a teaching assistant for an undergraduate statistics course while I was pursuing a bachelor's degree in mathematics education. This experience vastly supplemented my education. I perceived that many of my classmates had a gap in their education because our courses emphasized the mainstream calculus trajectory. I was an anomaly because I had initially majored in actuarial science (the study of statistics to quantify risk) before I changed my major to math education, so I took more courses in statistics than others in my program. Although each of us graduated with teaching licenses that endorsed us to teach advanced placement statistics, my classmates did not have many learning opportunities that would allow them to be successful teaching such a course. Several of my friends from the program openly admitted that they did not apply for math teacher positions if the job description involved teaching statistics. I hope to provide information in this study that may empower teacher educators with more tools to help their pre-service teachers learn about confidence intervals, and to consequently be able to teach confidence intervals. The purpose of my research is to contribute to the field of mathematics education by informing individuals who teach principles of confidence interval estimation, particularly to pre-service teachers, of developments in understanding as they progress from misconceptions toward expert conceptions. I intend to investigate the prevalence of misconceptions about confidence interval estimation in a group consisting of pre-service teachers who have already completed at least one statistics course. Further, I intend to inquire about the nature of experiences that served to perturb the student's thinking according to the students themselves.

CHAPTER 2: LITERATURE REVIEW AND THEORETICAL FRAMEWORK

Literature Review

The intent of my literature review is to provide detail concerning what is known about student conceptions related to confidence intervals since learning how students develop reasoning about variation is an important research topic (Garfield & Ben-Zvi, 2007). This will be accomplished by establishing what past researchers have concluded makes confidence intervals a difficult topic for students to understand. Next, I investigated what specific misconceptions about confidence intervals have been documented. After learning why researchers have claimed misconceptions may exist and also what those misconceptions are, I looked for studies about how misconceptions may evolve. Finally, since confidence interval computations rely extensively on theoretical sampling distributions that contain an infinite number of samples, I searched for studies regarding student conceptions of sampling distributions and of infinity.

First, researchers have claimed that although students can often compute confidence intervals, they have difficulty understanding what they mean (Garfield & Ben-Zvi, 2007). The justification given for students having difficulty understanding confidence intervals is that they often do not understand the concept of variation. Garfield and Ben-Zvi (2005) found that the practice of pointing the finger of blame for misconceptions about statistical concepts specifically at the concept of variation is a widespread practice. A justification for this conclusion notes that variability is a complex conception particularly because variability may sometimes be desired and of interest, and sometimes be considered error or noise (Konold & Pollatsek, 2002). It is also complex because of the interconnectedness of variability to concepts of distribution, center, sampling, and inference (Cobb, McClain, & Gravemeijer, 2003). An example of researchers blaming lack of understanding about confidence interval probabilities on a lack of understanding variation is Zhang and Stephens (2016). Zhang and Stephens (2016) said teachers who gave incorrect feedback about a situation were said to have not realized variation exists and the sample size will influence the variation. Gauvrit and Morsanyi (2014) used processes involving random variation to combat the equiprobability bias. The equiprobability bias is a belief that every process in which randomness is involved corresponds to a fair distribution, with equal probabilities for any possible outcome (Gauvrit & Morsanyi, 2014). A person's lack of understanding about confidence interval estimation in particular seemed to have been attributed to being a side effect of a lack of understanding of variation. This is often evident in the types of recommendations given to combat misunderstandings. For example, Blanco (2016) suggested that pre-service teachers' statistics education should include more opportunities to work with numerical data in order to explore variability. Thus, misconceptions about variation are generally considered to be the root of misconceptions about confidence intervals.

After finding compelling evidence that confidence intervals are difficult for students, I investigated what specific misconceptions students have about confidence intervals. A survey of the field performed by Castro-Sotos et al. (2007) proved in this venture. In their survey, meta-analytical methods were used to summarize many student misconceptions of statistics. To determine the important concepts to include in their review, the researchers included topics from handbooks of introductory courses on statistical inference (e.g., Healey, 2005; Moore & McCabe, 2006). Confidence interval estimation was deemed to be among the critical areas of understanding. Information about misconceptions in the critical areas of understanding were then summarized from publications gathered from ERIC, PsycINFO, and Web of Science databases (e.g. Haller & Krauss, 2002; Batanero et al., 2004; Chance et al., 2004; Vallecillos, 2000; delMas & Liu, 2005). Castro-Sotos et al. (2007) found a study by Fidler (2006) that outlined common misconceptions related to confidence interval estimation. Fidler (2006) detected six types of misconceptions associated with confidence intervals. I used only the five most common misconceptions as reported by Fidler (2006) to motivate my data collection decisions because Fidler observed the sixth misconception in less than five percent of the students studied. I was interested in common misconceptions because I wanted to design questions for pre-service teachers that would be likely to draw out misconceptions from many students. I wanted evidence of misconceptions to be present in pre-service teacher responses to the questions I designed so as to identify candidates for the case studies.

According to Fidler's study (2006), in which data was collected from 180 first and second year university students, the five most common misconceptions about confidence intervals are as follows. First, the most common misconception was that a 90% confidence interval is wider than a 95% confidence interval for the same data. The second most common misconception was that a confidence interval provides plausible values for the sample mean. The third most common misconception was that the width of a confidence interval is not affected by sample size. The fourth most common misconception was that the width of a confidence interval increases with sample size. The fifth most common misconception was that confidence intervals provide a range of individual scores within one standard deviation of the parameter. These five misconceptions are summarized in Table 1. As an attempt to increase the likelihood of observing the fifth most common misconception, I did not restrict this misconception to "one standard deviation" specifically. Rather, any instance in which a student confused a calculated confidence interval as being centered at the population mean regardless of whether their response conveys that they believe the range contains the middle one, two, or three standard deviations of the population data were considered to be a misconception of the fifth type.

Table 1. *Misconceptions of confidence intervals*

Description of a confidence interval	Percentage of students (n = 180)
A 90% confidence interval is wider than a 95% confidence interval	73
Confidence intervals give plausible values for the sample mean	38
The width of a confidence interval is not affected by sample size	29
The width of a confidence interval increases with sample size	20
Confidence intervals give a range of individual scores within one standard deviation of the parameter	11

In addition to the misconceptions given by Fidler (2006), another known misconception is that the confidence level gives the probability of a calculated confidence interval containing the parameter (Gilliland & Melfi, 2017). This conception is highly nuanced since it is only incorrect when applied to confidence intervals that have been calculated, but it is correct when applied to a collection of confidence intervals that have not yet been realized. Given a confidence level of 0.954 and assuming that the observations are normally distributed with mean μ and standard deviation σ, it is correct to say the probability that μ is within the interval created by adding and subtracting two standard deviations from the sampling distribution to a sample statistic is 0.954 for all μ. This statement is not correct if applied to a calculated (realized) confidence interval. In other words, $\Pr_{\mu}(\bar{x} \pm 2\sigma /\sqrt{n}$ captures $\mu) = 0.954$ for all μ is a true statement while $\Pr(46 < \mu < 54) = 0.954$ is a false statement. The statement is false because it implies that the parameter varies by being within the bounds some of the time and outside of the bounds at other times, but the paramenter is a fixed value. Once a confidence interval has been calculated, the probability that the parameter falls inside that interval is either zero or one regardless of the level of confidence because the parameter is either captured by the interval or it is not. For example, one might appropriately say that the chance of rain tomorrow is between 30 and 40 percent. However, they cannot appropriately say that the chance of rain yesterday is between 30 and 40 percent because it either rained yesterday (so the probability would be 1) or it did not rain yesterday (so the probability would be 0).

After finding these misconceptions, I conducted a search for more recent contributions to the field concerning how student misconceptions might develop into conceptions that are more in line with student thinking. I searched the Journal of Statistics Education database and found relevant articles by Pfaff and Weinberg (2009) and by Gilliland and Melfi (2017). Pfaff and Weinberg's study involved a hands-on activity aimed at improving student conceptions of confidence intervals, but it was unsuccessful. Their activity had focused on repeated sampling by having students repeatedly sample bingo chips from a bag and generate a prediction in the form of a confidence interval for the proportion of blue chips in the bag (Pfaff & Weinberg, 2009). Gilliland and Melfi (2017) also focused on repeated sampling when they noticed that the language often used in textbooks to define confidence intervals accepts *frequentist probability*. Frequentist probability is an interpretation of an event's probability as the limit of its relative frequency in a large number of trials, so this language links the definition of confidence intervals to a practice of taking many samples. Pfaff and Weinberg (2009) and Gilliland and Melfi (2017) asserted that although we know some factors that make confidence intervals difficult (such as variation), as a community of statistics educators we have not figured out how student conceptions might evolve into conceptions that are more in line with student thinking.

Finally, I investigated student conceptions about sampling distributions and infinity because confidence interval estimation relies on the concepts of sampling distributions that contain an infinite number of samples. For the benefit of orienting the reader, I will first provide an explanation of expert conceptions of confidence intervals and sampling distributions.

An expert conception of a confidence interval would consist, at the basic level, of two parts. First, it would be understood to be a range of values that could serve to approximate an unknown population parameter. In other words, the confidence interval could be an estimate for the set of values of the population parameter that were not rejected by the sample data (Gilliland & Melfi, 2017). Second, it would be understood that if many confidence intervals were created of that same sample size and same confidence level, "CL", essentially CL% of the confidence intervals would succeed in capturing the parameter. Thus, one could claim that there is a CL% probability that the confidence interval that was created is one of those that successfully captured the parameter.

Next, I will discuss an expert conception of a sampling distribution. By definition, a distribution must include all possible values (generally depicted along the x-axis) together with how often those values occur (generally depicted along the y-axis). A sampling distribution of $\bar{x}$ therefore includes all possible sample mean values that could be acquired by taking samples of size n and all their frequencies. There are three properties of sampling distributions, based on the Central Limit Theorem, that are often taught to prepare students for inferential statistics (e.g. Lane, 2014; Healey, 2005; De Veaux, Velleman, & Bock, 2009; Moore, McCabe, & Craig, 2009). The first property is that the mean of the sampling distribution of all $\bar{x}$'s is exactly equal to the population mean. This can be symbolically expressed as $\mu_{\bar{x}} = \mu$. The second property is that the standard deviation of the sampling distribution of $\bar{x}$ is equal to the standard deviation of the population divided by the square root of the sample size. Symbolically, this is represented as $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The third property is that the shape of the sampling distribution of $\bar{x}$ is normal if the population's shape is normal or that its shape becomes more normal as the sample size increases regardless of the population's shape because of the Central Limit Theorem.

Sampling distributions are fundamental for appropriate conceptualizations of the confidence interval. This is because a sampling distribution is the bridge between summary statistics and inferences about unknown parameters. Confidence intervals are inferences about unknown parameter values based on sample statistics, and sampling distributions are used in theory to arrive at those inferences with a certain level of confidence. The theoretical distribution of sample means about the population mean lets us infer the probability of a random sample mean being within a given range of the parameter. Without building a sampling distribution by collecting an infinite number of sample means, the properties of the sampling distribution and the inferences that rely on it are still valid. Hence, confidence intervals are built based on one, single sample so long as the sample is large and random.

I include research about student conceptions of sampling distributions in this section because confidence interval estimation relies on using sampling distributions, so conceptions about confidence intervals are destined to be tied to conceptions about sampling distributions. Research showed that the concept of sampling distributions is prone to misconceptions. For example, Watkins, Bargagliotti, and Franklin (2014) found that using simulations to demonstrate sampling distributions may cause students to confuse properties of the sampling distribution. Specifically, when students observe simulations of sampling distributions with increasing sample sizes, they may believe that the mean of the sampling distribution approaches the mean of the population and that the standard deviation of the sampling distribution approaches the standard deviation of the population divided by $\sqrt{n}$ (Watkins et al., 2014). This misconception is reasonable since students can intuitively understand that bigger is better, so they think bigger samples give better approximations for the mean and standard deviation (Watkins et al., 2014). Although logical, this thinking is incorrect since for any given sample size n, the mean of the sampling distribution is exactly equal to the mean of the population and the standard deviation of the sampling distribution is exactly equal to the population standard deviation divided by √n. The "approaches" language could be appropriate if the sampling distribution did not include all possible sample means. However, by definition, the sampling distribution does include every possible sample mean, so the properties of the sampling distribution that compare the means and the standard deviations of sampling distributions and populations are exact. Specifically, $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, and the misconception is that $\mu_{\bar{x}} \approx \mu$ and $\sigma_{\bar{x}} \approx \sigma/\sqrt{n}$.

I also include research about student conceptions of infinity because sampling distributions deal inherently with the idea of repeating a process of collecting an infinite number of samples and creating a resulting object under the limit, they deal implicitly with the ideas of infinity. Similar to sampling distribution, infinity is abstract in nature and difficult to link to real-life experiences. Understanding infinity is difficult and is dependent on one's ability to visualize mentally (Kolar & Čadež, 2012). Piaget and Inhelder (1956) studied student responses to what the final element would be if a geometric figure were continuously divided by two mentally. They learned that children's ability to visualize the division of a geometric figure into smaller parts is limited to a finite number of iterations (Piaget & Inhelder, 1956). The number of samples involved in sampling distributions may likewise be beyond what students can conceptualize.

Potential and actual infinity are two terms used for how a student conceptualizes infinity. Potential infinity is "conceptualized through iterative processes that may be repeated on and on, and the completion of this process is external to the act" (Hannula et al., 2006, p.4). This may be compared to the counting process which we know can go on and on without a boundary. "Actual infinity is the end of the infinite process. Actual infinity cannot be conceptualized through potential infinity process alone, but it requires the conceptualization of an end point of the process” (Hannula et al., 2006, p.5). These terms were useful to my study because they helped me classify student descriptions of sampling distributions as processes or objects.

I investigated why focusing on repeated sampling may not have been successful in helping improve student conceptions using literature about common misconceptions of confidence intervals, sampling distributions, and infinity. I wanted to know if misconceptions about confidence intervals were associated with misconceptions about variation for the students in my study. If misconceptions about variation were not associated with misconceptions about confidence intervals, I wanted to learn what other misconceptions exist that might be problematic and what the sources of those misconceptions might be.

Theoretical Framework

Knowledge in Pieces

The basic stance that underlies my study is Knowledge in Pieces (KiP) pioneered by Andrea diSessa. In this perspective, there are two different types of concepts: categories and obtaining information. A category conception's primary cognitive purpose is to differentiate whether something is or is not part of that category (diSessa & Sherin, 1998). For example, we may use our conception of a bird to determine whether or not something is a bird. Other possible category examples include cars, flowers, and mothers. Coming to understand one of these categories is to come to understand what counts as being that thing and maybe having a sense of what examples are the best representatives of that category and what variations can exist within that category. However, this notion of concept as a category does not adequately account for all the things we call "concepts," so diSessa and Sherin (1998) promoted a second type of concept with the primary cognitive purpose of obtaining information. This specific type of concept that deals with obtaining information is called a *coordination class.* For example, consider the concept "velocity." The point of this concept is not determining if something *is* a velocity, but deals with obtaining information like how fast and in what direction an object is traveling. Other examples of coordination classes include (a) force because it deals with information like what the magnitude and direction are, (b) area because it deals with information like how much space exists in an enclosure, and (c) number because it deals with information concerning how much of a quantity exists. Confidence intervals could also be viewed as an example of a coordination class because one primarily uses a concept of confidence interval to gather information about where a parameter may be rather than to determine whether something is or is not a confidence interval.

Coordination classes deal with obtaining information through readout strategies and causal nets. A *readout strategy* is the way a person interprets or internalizes external information (Levrini & diSessa, 2008). For example, one may read the math symbol μ as the mean of a population, or the symbol expression (a,b) as representing an interval of numbers starting at *a* and ending at *b*. A *causal net* is a system of how an individual's knowledge elements activate each other (Levrini & diSessa, 2008). By activate, I mean that one knowledge element triggers the individual to think of a subsequent knowledge element. For example, the math symbol μ may trigger the activation of population standard deviations or of sample means. The context for which a causal net strategy is applied is called the *concept projection*.

Knowledge is said to be in "pieces" because rather than necessarily existing as a set of integrated structures and theories, it is believed to exist in a large number of individual pieces (diSessa, 1988). Clusters of these individual pieces can be activated and used in different combinations, without necessarily having every possible knowledge element associated with that concept being activated. Among these individual pieces are *p-prims*, which are the most basic elements of knowledge, defined as "simple abstractions from common experiences that are taken as relatively primitive" (diSessa, 1988, p. 52). Thus, causal net elements, or knowledge pieces, might consist of either basic p-prims or larger grain-sized knowledge resources that have been constructed from p-prims or formal education. An example of a p-prim causal net element is "being closer equates to being more intense," such as being closer to a fire makes it hotter or being closer to a lightbulb makes it brighter. This is a "simple abstraction from common experience." An example of a p-prim causal net element for confidence intervals is that more confidence equates to more certainty, in that increasing the confidence level increases the likelihood of being "right." The idea that confidence activates knowledge pieces like certainty, or perhaps accuracy, is an idea that is naturally created as people experience life. Of course, knowledge elements in a coordination class can be more than p-prims, such as knowing that a “sample” is a subset of the population. This is more than a simple abstraction, and represents a constructed piece of knowledge, though it is still a single element that can be used in a causal net.

Defining Misconception

In this paper, I use the term “conception” to mean a person’s overall readouts and causal nets associated with the construct “confidence interval.” I call them “*expert conceptions”* when they have developed into normative conceptions generally accepted among experts, such as those belonging to the statistical community. Since even experts can still be mistaken, an *expert conception* is not a conception that is perfectly developed and flawlessly applied, but a conception that has reached a stage of relative mastery as accepted among experts. Oftentimes, researchers refer to instances in which conceptions are in conflict with expert conceptions as *misconceptions* (Confrey, 1990). Sometimes the traditional use of the word “misconception” assumes a stable cognitive entity that a student has constructed in association with that concept. However, as diSessa (1988) and Smith et al. (1994) have argued, misconceptions might be due to the activation of a useful knowledge element in the wrong context.

The use of the term *misconceptions* in this study will refer to an issue with readout strategies or casual nets that, for my purposes, can happen in one of two ways. First, one source of misconceptions may be missing or incorrect readout elements or causal nets. This is closer to the traditional usage of the word “misconception,” though it deals more subtly with individual knowledge elements rather than an entire concept. Related misconceptions could be incorrect interpretations of μ, p, margin of error, the symbolic parenthesis around confidence intervals, $\bar{x}$, p-hat, and so forth. For example, a student misconception may be that p-hat is the probability of observing a statistic as extreme or more extreme than was observed assuming that the null hypothesis is true. However, this definition is suitable for a p-value, not for p-hat which is a sample proportion, so this would be an example of an incorrect readout element. The second, and possibly more important, source of misconceptions may be inappropriate links between two otherwise useful knowledge elements. An example of this could include a student misconception that changing the size of a sample has no impact on the width of a confidence interval. This student may be activating a p-prim element that changing the size of the sample does not impact the size of the population or the respective parameters, and then inappropriately believing that it therefore has no impact on the resulting confidence interval. The idea of sample size not impacting the population size is *not* incorrect, and does not need to be "confronted and replaced" as is traditionally believed necessary for misconceptions. It is simply the wrong inference to make in this specific context of confidence intervals.

Misconceptions are not viewed in this paper as adversarial. Rather than being in conflict with expert conceptions, misconceptions will be a considered a common and even necessary stage preceding the level of expert conceptions (Smith et al., 1994). This view of misconceptions is consistent with a KiP perspective because as students construct knowledge, their understanding may initially take the form of a misconception. Misconceptions are developed when one draws on an incorrect assumption or knowledge resource, but not because the individual is thinking in illogical or unreasonable ways (Smith et al., 1994). Misconceptions are not mistakes that impede learning. They are evidence of the very act of learning itself. For example, my four-year-old daughter draws the letter E using more than three horizontal lines. By my definition, her E indicates that she has a misconception because her understanding has not developed into an expert conception as accepted by people proficient in writing letters of the English Alphabet. However, she can use her current misconception to progress toward an expert conception by eventually recognizing that only three horizontal lines are necessary. This means that although she is mistaken, her misconception can still be productive. This perspective of misconceptions being potentially productive was explained by the Smith et al. (1994) who stated:

> [The principles] conceptualize knowledge not in terms of the presence or absence of single elements (e.g., F = ma or conversion to common denominator) but as knowledge systems composed of many interrelated elements that can change in complex ways. This knowledge system framework makes it easier to understand how novice conceptions can play productive roles in evolving expertise, despite their flaws and limitations (p. 117).

Further, diSessa discussed that an attack on every misconception students possess is not only hopeless, but is akin unto throwing the baby out with the bath water. He stated "the only material we have to develop scientific understanding in our students' heads is precisely those [misconceptions]" (diSessa, 1988, p. 51). The KiP perspective guided how I thought about and understood the roles of confidence interval misconceptions as a teacher gains the expertise needed to teach this concept to their own students. In particular, it shows that, rather than trying to avoid or prevent misconceptions in students, it might be more helpful to understand how these misconceptions could be an integral part of the story of how teachers develop their thinking of statistics. All learning involves the interpretation of situations, including classroom instruction, through the perspective of the learner's existing knowledge (Resnick, 1987; von Glaserfeld, 1987). Students do not come in the classroom as blank slates, but rather, they bring with them their prior learning. Oftentimes, this means that students will have misconceptions about certain principles from their interaction with the physical and social world. Educators should not engage themselves in a crusade to destroy misconceptions exhibited by their students. Instead, educators could seek to learn about students' existing, underlying thinking that is evidenced by the misconceptions. It is my hope that this approach will lead students to develop expert conceptions by seeing their former conceptions in a more nuanced way, not by seeing their former thinking as wholly wrong, and that educators may come to understand the root of the thinking that leads students to misconceptions.

There is not a clear boundary between misconceptions and expert conceptions since once an expert conception is attained one may still later use reasoning appearing to involve misconceptions. This could, again, be due to drawing on a useful knowledge resource in the wrong context. An expert would not discard useful knowledge resources that they may productively use in some situations, but they may accidentally apply it in the wrong context. For example, the knowledge element "closer means more intense" may be used to correctly describe heat flow and transfer, but it could accidentally be used incorrectly to justify why the Northern Hemisphere is warmer in the summer, despite the actual reason being different. Further, dilineation between misconceptions and expert conceptions can be unclear because a student's expert conception may only be manifest in certain contexts. For example, when cognitive load increases "previously held misconceptions or erroneous approaches may be brought to bear, reinforcing knowledge that is counter to the material they are trying to learn" (Feldon, 2010, p. 18). Thus, when I report evidence of a student holding an expert conception, I do not believe that the student will exhibit the expert conception in every situation.

Conceptions are not held stable, and confidence interval conceptions are especially subject to change throughout one's education. There is often a battle with intuition and academic experiences. Morsanyi, Primi, Chiesi, and Handley found that furthering one's education does not always correspond to a decrease in misconceptions regarding statistical concepts (2009). In fact, misconceptions such as the equiprobability bias increase with probability education (Morsanyi et al., 2009; Gauvrit & Morsanyi, 2014).

Student Noticing

Due to my interest in studying possible sources for student misconceptions, it is necessary to include a framework that deals with the relationship between classroom experiences and readout strategies or causal nets developed by students. Student interpretations and inferences based on their experiences in the classroom may sometimes unintentionally lead students to develop misconceptions (Jones, Lim, & Chandler, 2017). This belief is captured through the notion of *student noticing* (Lobato, Rhodehamel, & Hohensee, 2012). Student noticing is defined as "the selection of certain information in the presence of competing sources of information" (Lobato, Rhodehamel, & Hohensee, 2012, p. 438). The visual, verbal, or conceptual objects that a student pays attention to are called *centers of focus* (Lobato et. al., 2012).

Research Questions

In my study, I examined a small sample of pre-service teachers who were enrolled at the time in a mathematics education course entitled, "The Teaching and Learning of Statistics and Probability." The focus of this class was on developing the pre-service teachers' understandings of statistics and probability concepts, and I wanted to investigate what misconceptions they may have held about confidence intervals. In particular, my study is centered on the following three research questions:

1. What misconceptions about confidence intervals exist for the group of pre-service teachers in this study?

2. What in-class experiences may have led to those misconceptions?

3. What interpretations or inferences led to changes in their conceptions?

CHAPTER 3: METHODOLOGY

Setting and Participants

The participants of this study came from a large university in the United States and were enrolled in a mathematics education course for pre-service secondary teachers focused on the teaching and learning of statistics and probability, entitled “The Teaching and Learning of Statistics and Probability.” The course is required for students pursuing a bachelor’s degree in mathematics education at this institution in response to common core standards and because many graduates from the program admitted in exit interviews that they felt inadequate to teach statistics. The learning outcomes associated with the course focus on developing student understanding of study design, summarizing and representing data, drawing conclusions from data, probability, and research on students’ understanding of statistics and probability. As a prerequisite, students had taken at least one course in basic statistics and at least one course in mathematics education prior to taking this course. Consistent with general demographics of the institution and of the mathematics education major, participants were anticipated to be mostly White females in their early twenties. There were twenty-five students enrolled in The Teaching and Learning of Statistics and Probability course who chose to participate in the fall of 2017 when data collection took place. Data was collected from 25 students in the class who consented to be involved in the study. Based on that data, five students were invited to participate in interviews. In this paper, I give the five students the pseudonyms Ethan, Danielle, Tiana, Anna, and Corinne. All of these students were between the ages of 18 and 25. Ethan is a White male. He did not take statistics in high school. His first statistics course was an introductory statistics course at the university, which I call “Introduction to Statistics.” He took the course two years before he was interviewed for this study, and he described it as the “easiest class that [he] had taken in [his] college career." He said he used a lot of statistics in physics courses associated with his double major. Danielle is a White female. She took also took Introduction to Statistics and was confident that she had one of the highest percentages in the class. Tiana is a non-White female who took statistics in high school, but she said her teacher "clearly did not understand statistics well so the course was not the best." She took Introduction to Statistics in college and thought it was "fairly easy." Anna is a White female. She did not have a statistics course in high school, and she took Introduction to Statistics a year and a half prior to being interviewed for this study. She expressed that she felt very competent doing computations in Introduction to Statistics and received a good grade, but that she did not know why the formulas work. Corinne is a White female. She did not take statistics in high school and she did not take Introduction to Statistics. She said the only statistics class she had before The Teaching and Learning of Statistics and Probability course was a statistics course for engineers in college that taught her to use a software program to analyze data. She said that she was not taught theorems about statistics, but she had a lot of practice with computations. The participants' demographics are summarized in Table 2.

Table 2. *Participant demographics*

	Gender	Statistics in High School	Statistics in College	Successful in Previous Statistics Course(s)
Ethan	Male	No	Introduction to Statistics	Yes
Danielle	Female	No	Introduction to Statistics	Yes
Tiana	Female	Yes	Introduction to Statistics	Yes
Anna	Female	No	Introduction to Statistics	Yes
Corinne	Female	No	Statistics for Engineers	Yes

Data Collection

My data collection consisted of three phases, including a pretest, first interview, and second interview. Semi-structured interviews were selected as a primary source of data collection, although responses to class activities and homework assignments were also reviewed as a way to inform the direction for the case study interviews. The protocols for each of the two interviews allowed for ample flexibility in student responses and subsequent interviewer questions. The following is a description of each phase of data collection.

First, near the beginning of the semester each of the students in The Teaching and Learning of Statistics and Probability course who consented to participate completed a pretest as part of their normal coursework. The pretest was designed to investigate the prevalence of the five common misconceptions found by Fidler (2006), described in the Literature Review section. The pretest allowed me to gather evidence of students' existing conceptions about what confidence intervals are and what the purpose of an interval is so that I had an entry data point for the students. Students were encouraged to put their best guess for each answer. They were also directed to indicate their level of conviction for each of their responses on the pretest on a scale from 1 to 5. Students were not told their score on the pretest. The pretest questions are given in Appendix A.

The purpose of the pretest was to identify good candidates for the interviews. I was looking for students who exhibited one or more of the common misconceptions and who could articulate justifications for their responses. I also hoped to gather a variety of misconceptions across the students who were selected.

Following the pretest, I sent private e-mails to seven of the pre-service teachers to participate in interviews with the intention of completing both rounds of interviews with five of these students. The purpose in selecting more than five students was to have a contingency plan in place for any students who dropped out during the case study and allow me to choose the most interesting student responses for my final analysis. I only got responses from five of the students. Four of the five students completed both interviews and one student (Tiana) completed just the first interview.

The second phase of data collection was the first round of interviews. The purpose of the first interview was to better understand the students' prior experience with confidence intervals, their existing conceptions about confidence intervals, and the extent to which they were convinced by their conceptions Students were given copies of their pretest without any feedback. They were asked to go through each of their responses and clarify their justifications. In this process students sometimes modified their original responses, and I questioned them about why they thought their first answer was incorrect or insufficient. The first interviews lasted 30 minutes each. The interviews were conducted on campus and were audio recorded and transcribed. The specific protocol can be found in the Appendix B.

Between the two interviews I attended and recorded information from lectures in The Teaching and Learning of Statistics and Probability course. I was using a non-intrusive system built directly into the classroom that allowed me to sit at the edge of the classroom and operate cameras that were mounted in the ceiling. The purpose of these recordings was to compare potential accounts of student noticing with the pedagogical practices taking place in the classroom.

Third, after the pre-service teachers had class sessions concerning confidence intervals I held another round of interviews. The second round of interviews was completed with Danielle, Anna, Ethan, and Corinne, but Tiana was unable to meet with me. The second interviews lasted one hour each and were audio recorded and transcribed. The purpose of the interview was to learn what the students' existing conceptions were after having instruction in The Teaching and Learning of Statistics and Probability course and to document their experience. When misconceptions that were present in the initial interviews were still evident, follow up questions were asked to gather data that could be analyzed to identify the student's concept projection. When misconceptions present in the initial interviews were not evident, follow up questions were asked to gather data that could be analyzed to determine why the student changed their thinking. All of the students were asked to walk me through a hypothetical process of calculating a confidence interval beginning at the data collection stage. The parameter of interest was dependent on contexts that the students selected themselves. In both interviews I asked students to state what information confidence intervals provide and to define what a sampling distribution is. I took on the role of a student as I listened to their explanations and as I asked them questions. My questions were not designed to get any of the students to correct their misconceptions though there were many instances of students throughout the interviews revising their previous answers as I tried to learn about their conceptions. To learn about their conceptions, my questions often pressed for the students to explain why their claim was true, how something works, and whether they believed their claim would hold under different conditions.

The first interviews took place in the middle of the semester and the second interviews took place at the end of the semester. The semester began September 5th and the last day of final exams was December 21st. The first interviews were completed between mid-October and early November, and the second interviews were completed between late November and mid-December as summarized in Table 3.

Table 3. *Participant interview dates*

Interview 1 Dates		Interview 2 Dates	
October 16	Ethan	November 30	Ethan
October 20	Corinne	December 11	Corinne
October 20	Danielle	December 11	Anna
October 20	Anna	December 15	Danielle
November 2	Tiana	Not Applicable	Tiana

Data Analysis

Coding Pretests

For the pretest, the coding was based on a continuum scale since misconceptions and expert conceptions are not separated by a clear boundary. Each conception could be used productively by the students working toward increasingly expert conceptions, so the spectrum of conceptual understanding was considered in degrees of proficiency. The student responses were coded as showing evidence for conceptions that are minimally proficient, emerging proficient, proficient, or highly proficient as in Table 4. In Level 1, the student response is incorrect and has no justification. In Level 2, the student response is incorrect, but has a justification. In Level 3, the student response is correct, but has an inappropriate justification. In Level 4, the student response is correct, and the justification is appropriate. For the purposes of this particular study, I considered Levels 1 and 2 to be "misconception" levels while I considered levels 3 and 4 to be "expert conception" levels.

Table 4. *Levels of proficiency for coding student responses*

Level 4: Highly Proficient	The response is correct and the justification is sound
Level 3: Proficient	The response is correct but the justification is faulty
Level 2: Emerging Proficient	The response is incorrect but the student was able to provide a justification
Level 1: Minimally Proficient	The response is incorrect and the student did not provide a justification

Following are examples of student responses to a pilot survey question. An example is provided for each of the four coding levels. The students were asked to imagine that there were 100 different researchers who each took a random sample of size 50 from the same population. Given this scenario, the students were asked to explain their reasoning about the approximate number of resulting intervals that would not have captured the population mean if each researcher constructed a 95% confidence interval for the mean.

Example of a Level 4 Response: "Approximately 5%. 95% confidence interval means that if this process is used repeatedly, the true mean will lie in the interval 95% of the time."

This response received a code of *highly proficient* because the student provided an answer that was consistent with expert knowledge (approximately 5%) and she defended her answer by incorporating a sound, relevant description of what it means to use a process with a specific level of confidence.

Example of a Level 3 Response: “5% this is the interval allowed for faultiness so it [is] constructed at a 95% confidence interval, [only] 5% would not capture the population mean.”

This response received a code of *proficient* because the student provided the correct answer of 5% which is in line with expert thinking, but the justification was problematic. Specifically, the justification was problematic since their use of the term “interval” is inconsistent with the meaning of confidence interval. Further, her justification repeats 95% confidence as meaning 5% faulty, which seemed to be making the statement that her answer was true by tautology.

Example of a Level 2 Response: “I thought that the confidence interval was made around a mean, so they should all include the mean.”

This response received a code of *emerging proficient* because the student answered incorrectly, but provided a justification for the response. The misconception evident is that a confidence interval gives plausible values for the sample mean. The justification given was because confidence intervals are “made around a mean” but it is reasonable to assume the student confused the sample mean and the population mean. Note that confidence intervals are in fact made around a mean, so this misconception can still be a productive conception.

Example of a Level 1 Response: “These confidence intervals will not capture the population mean because confidence. Looking at this question I now see that I don’t actually remember what confidence intervals are and I think I’m getting it confused with confidence levels, but I tried.”

This response received a code of *minimally proficient* because the student did not provide the answer of 5% that is accepted among experts. Further, there was not a justification provided as to why “confidence” makes it so that the confidence intervals would not capture the population mean

After each of the pretests was coded by the level of proficiency, I grouped individual responses into categories based on the type of misconception that was evident. Responses without evidence of misconceptions were not considered. These categories were based on the misconceptions from Fidler (2006). I did not gather any strong evidence for misconceptions about the effect of confidence level on the width of the interval, so I had four groups that I was making selections from. Within each of the four groups I sought out responses with well-articulated justifications for the misconceptions because I believed they were associated with students with whom I would be able to conduct the most productive interviews. The well-articulated justifications for misconceptions were associated with the coding Level 2 since Levels 2 was used when misconceptions were present and students attempted to provide a justification for their response.

Students were also prompted to rate their level of conviction for each of their responses on the pretest. The ratings went from 1 to 5 and the students circled an option with 1 being associated with low levels of conviction and 5 being high levels of conviction about the validity of their response. I will report the student’s rating of their conviction for the misconception that they exhibited which caused me to invite them to participate in the study.

Coding Interviews

In the interview transcripts I looked for readouts (any time the student interpreted anything) and causal nets (knowledge pieces and times when they make inferences between those knowledge pieces). This process is described in greater detail in the next section. Next, I classified each of the readouts and causal net elements using a secondary code based on whether the student's statement aligned with an expert conception or a misconception. My final sorting was to categorize the student conceptions according to a theme. Themes were statements summarizing the misconception or expert conception manifest in the student's statement. I had pre-existing codes that I anticipated to be themes based on the common misconceptions from Fidler (2006) as shown in Table 5. I wrote expert conceptions that corresponded to each of the common misconceptions according to Fidler (2006) which were also included in the pre-existing codes (also shown in Table 5). I allowed for open coding since it was possible for a student to have a conception that was not one of the pre-existing themes.

Table 5. *Pre-existing codes*

Themes Related to Expert Conceptions about Confidence Intervals	Themes Related to Misconceptions about Confidence Intervals
Decreasing level of confidence for the same data results in a narrower interval	A 90% confidence interval is wider than a 95% confidence interval (for the same data)
A confidence interval gives plausible values for the population mean	A confidence interval gives plausible values for the sample mean
Increasing sample size results in a narrower interval	The width of a confidence interval is not affected by sample size
	The width of a confidence interval increases with sample size
The confidence interval could only reflect the middle __% of the data if the sample mean happened to be exactly the same as the population mean for Normally distributed data.	A confidence interval gives a range of individual scores within ___ standard deviation(s)

I wrote the associated expert conception themes using the condition that they give rise to an idea an individual may be able to perceive after the acquisition of the expert conception that they would have not been able to perceive without it. With the first theme, the student can see the relationship between level of confidence and the width of a confidence interval. They can understand that to be more confident in the results one must produce a wider interval to include more options for the unknown parameter estimate. With the second theme, the student can see that the purpose of a confidence interval is to estimate a parameter. They can understand that parameters generally remain unknown, but samples come from the distribution centered at the parameter of interest so the sample statistics provide an approximation of the parameter. With the third theme, the student can see the relationship between sample size and the width of a confidence interval. They can understand by increasing sample size the spread of the sampling distribution decreases so its standard deviation is smaller resulting in smaller margins of error and confidence intervals. With the fourth theme, the student can see the role of variability of sample means about an unknown population mean. They can understand that a sample may come from any part of the distribution so adding and subtracting the margin of error may result in an interval capturing the parameter on one end while including values that are scarcely overlapping with the distribution at all on the other end.

I narrowed down the themes that resulted from pre-existing codes and from open codes according to three criterion. First, I reported on the themes that were most prominent because I wanted to have enough data to support my assumptions about how the students were thinking. I decided that for a theme to be prominent, it had to be manifest by more than one student during the interviews. Second, I reported on themes that involved misconceptions since the focus of my study was to learn about student misconceptions and their possible sources. Third, I reported on any themes involving student conceptions of variation since existing literature generally claims that misconceptions about confidence intervals are rooted in misconceptions about variability.

I did further analysis on the themes that were established at the end of the aforementioned process. The further analysis was separated into three parts each intended to provide answers to my research questions. First, I looked in the transcripts for specific evidence that a student may have had a misconception regarding the theme. I did this by identifying inappropriate readouts or causal net elements. Second, I looked for connections to classroom experience that may have unintentionally influenced student misconceptions. I did this by looking for instances in which the students stated pedagogical practices, activities, or images that led to their misconceptions. Third, I reviewed the transcripts for instances in which a student was able to resolve a misconception, at least in part, and noted the inferences that led to the resolution**.**

Identifying Readouts and Causal Net Elements

To identify a readout strategy I looked for how the student interprets any external thing. For example, what a student says a symbol such as μ means will be considered a readout strategy. For causal nets, I identified the knowledge pieces that are brought up by the students themselves, that was not explicitly presented externally to them. I used the order of knowledge activation to infer what knowledge piece activated the next piece and so on. Students may appear to develop expectations about consequences of changing factors such as confidence level or sample size and compare their empirical observations against those expectations. I considered students' expectations to be causal net elements (as in Thaden-Koch et al., 2004). Every time I encountered a misconception in my data, I considered whether it is made up of individually correct ideas that are just put together inappropriately. I used the data compiled from students in the case studies to get a snap shot of how portions of their coordination classes behave. For example, if a student says that $27<\mu<29$ means that the population mean is between 27 and 29, then they have readout μ as the population parameter. If a situation changes n from 100 to 200 and the student says that the interval should get bigger because there are more people in the sample (a misconception), then they have readout n as the number of people in the sample and readout n of 100 then 200 as an increase in the number of people sampled. An applicable piece of the casual net is then that increasing the number of people in the sample activated the idea that more space would be required, and a subsequent causal net link is that if more space was required then the interval would become larger. In real life, more people do require more space, so this misconception is the connection of two otherwise correct ideas but that are problematically connected in this context.

I now illustrate the data analysis through examples in my pilot data. When describing that increasing confidence level decreases the width of the interval (a misconception) a student said, "I like to look at [confidence level] as like a sniper and a shotgun and the more precise that you are [by increasing level of confidence] the narrower your confidence interval is and the more sure you are." Here, two causal net elements are present because confidence level activated precision, and precision activated a sniper metaphor. In real life, since snipers do need to hone in on their target, that is an appropriate connection to have linked with precision. However, it became problematic in the mathematical context because more "precision" mathematically means increasing one's margin of error. At the end of the semester this same student reflected on that metaphor and said, "But that's not right, because decreasing the width of the confidence interval means that we're lowering the level of confidence because there are less possible values that could be there and so we are less sure that the true proportion lies within our confidence interval." She provided an alternative example stating, "It's like predicting the weather and being like there is a 99% chance that it will be between negative 100 degrees and like 5,000 degrees, and like you can be more confident that the weather is going to be there because you have a bigger interval. So the bigger the range, the more confident we are because there are more numbers for it to be." From an analysis of this data I can conclude that when this student was exhibiting a misconception, it was because she was applying an inappropriate metaphor, specifically with a sniper and a shotgun. This metaphor is still appropriate to other contexts, so it can still exist within her causal net once she has obtained an expert conception, but the connection between the sniper metaphor and the level of confidence was no longer where her readout strategies were directed. Level of confidence was introduced to her which induced a response in the form of searching for a causal net element. In her misconception model, this led her to think of precision which led her to think of the sniper metaphor. The sniper metaphor tended toward a conclusion that higher levels of confidence were associated with narrower confidence intervals. To address that misconception, she worked back to her concept of precision and confidence and realized that when one increases a level of confidence, they are going to have to cover more of their bases as she described by giving an example of a large range of possible temperatures.

CHAPTER 4: RESULTS

In this section I first explain why each student was invited to participate based on results from their pretest responses. Next, I describe the themes that surfaced from coding the interview transcripts. Following a description of the themes, I include results from my analysis of the misconceptions.

Participant Pretest Data

Following is an explanation of why each of the five students who were involved in the case studies were invited to participate. Seven students in total were invited to participate, but two students did not respond. Each of the misconceptions described below was coded as a Level 2: Emerging Proficient. That meant that the response was incorrect, but that the student provided a justification for the incorrect response. The level of conviction that the student chose associated with the response will also be reported.

Tiana was selected because she exhibited the misconception that confidence levels tell us the percent of the population data the lies inside the confidence interval. In the two questions where Tiana's misconception was evident she selected a 4 and a 5 on a scale of 1 to 5 where higher numbers were associated with higher levels of conviction that her answer was correct.

Ethan was selected because one of his responses indicated that he believed a confidence interval provides information about probabilities associated with a single individual rather than with the mean of the population. He had selected a 2 for his level of conviction associated with that response.

Anna was selected because she did not recognize that changing the sample size impacts the width of a confidence interval. She had selected a 3 as her level of conviction for that response. Further, Anna wrote two responses that indicated the belief that a confidence level tells us the percent of the time that the parameter will lie inside of a calculated confidence interval. For example, she wrote, “[the confidence level] tells us that 90% of the time our mean will lie in that range from 6.8 to 8.1.”

Corinne was selected because of responses that indicated two misconceptions. She, like Anna, expressed a belief that a confidence level tells us the percent of the time that the parameter will lie inside of a calculated confidence interval. Her level of conviction was just 1 on that response. Corinne also showed evidence of a misconception about the effect of sample size on the interval because she believed that sampling 500 or 100 individuals would result in having “about the same” upper and lower boundaries for the confidence interval. She seemed to believe that any difference in the interval’s bounds would be due to potential difference in the spread of the data for n = 500 and n = 100, but she did not expect that there would be much of a difference in the spread between the two samples. On the response Corinne selected 4 as her level of conviction.

Danielle was selected because her responses included many additions (shown by ^ marks drawn on her paper) and many revised attempts (shown by lines striking through portions of her sentences). For example, in responding to the prompt ‘What does a confidence interval mean?’ she wrote, “A confidence interval means that by repeated sampling with the same number of people sampled (^ and level of confidence) that (extensive scribbling) we will get intervals that (^ contain) the true mean (^ or proportion) for (scribbled word) the population (crossed out ‘is found in’).” This response did not incorporate that the number of intervals containing the parameter would be reflective of the chosen level of confidence, but her conception was not necessarily wrong. She reported 4’s and 5’s for her level of conviction. Danielle’s responses did not provide any clear indication of misconceptions, but I anticipated that she might be able to explain helpful metaphors that caused her to revise her reasoning since she was so actively amending her response.

Interview Themes

I identified four themes associated with misconceptions and one theme associated with an expert conception during my data analysis. The results of the misconception analysis are separated into two sections. The first section pertains to the two misconceptions in which properties of individual data points were mistakenly regarded as properties of confidence intervals. These misconceptions were anticipated prior to data collection because they were reported in the literature. The presence of these misconceptions was often revealed when students stated their beliefs about what information confidence intervals provide. The second section pertains to the two misconceptions regarding the conceptual role of repeated sampling when constructing a confidence interval. These misconceptions were not anticipated prior to data collection as they were not previously reported in the literature. The presence of these misconceptions was often revealed when students attempted to provide a conceptual justification for an interval's level of confidence.

Following is a brief description of each of the themes. The first misconception theme was that confidence intervals can be used to make predictions for individuals or that the mode was the same as the mean or median for a data set that was not necessarily normally distributed. Mode, mean, and median were put together into one theme since those values are approximately equal to one another when data is normally distributed. The second misconception theme was that the percentage given by a confidence level tells us the percent of the population data encompassed by a confidence interval. The third misconception theme was that we cannot work with an actual sampling distribution. The fourth misconception theme was that we must take multiple samples to compute a confidence interval. The fifth theme concerned variation. The participants in this study each showed evidence of possessing expert conceptions regarding the variation of a sample mean about the population parameter in the context of a sampling distribution.

Table 6 indicates which of the four misconception themes were observed in interviews with each of the students. I did not structure the interviews to gather data from every student at each interview concerning each of the misconception themes. Because I did not gather data about the students' conceptions regarding each theme, not observing a misconception during an interview may simply mean that the interview did not address that specific misconception with that student. For this reason, misconceptions may have been more prevalent than indicated in Table 6.

Table 6. *Misconceptions observed in interviews with participants*

	Observed Misconception #1 in Interviews	Observed Misconception #2 in Interviews	Observed Misconception #3 in Interviews	Observed Misconception #4 in Interviews
Tiana	No	Yes	Yes	Yes
Danielle	No	No	Yes	Yes
Anna	Yes	No	Yes	Yes
Corinne	No	Yes	Yes	Yes
Ethan	Yes	No	Yes	Yes

Another note about the following report is that it is geared toward considering confidence intervals built using univariate, quantitative data only. Generally, the participating students and I referred to $\bar{x}$ as the sample statistic without mentioning other possibilities such as sample proportion or sample slope. The initial pretest that I provided to students included a scenario with quantitative data, but I made no attempt to restrict our discussions to exclude qualitative data or bivariate data. The practice of discussing confidence interval topics in terms of μ and $\bar{x}$ was consistently perpetuated by each of the students in the study, so the case of a confidence interval built around a sample mean with the sampling distribution specifically being a sampling distribution of $\bar{x}$ will be the only case addressed in this study.

Each of the four sections about misconceptions is divided into three parts. Following a brief introduction of the misconception, the first part provides evidence of the misconception. In this section I use excerpts from interviews as examples of times I identify the presence of the misconception. The second part provides connections to classroom experience. For the second part, I use excerpts in which students might have used the curriculum or activities from a statistics class to support their misconception in a way that educators would not have intended. The third part describes how the students resolved the misconceptions. In the third section, I use excerpts to show how certain questions or situations that the students considered during the interview may have served to move the students' thinking toward more expert conceptions.

Misconception 1: A Confidence Interval Allows Us To Make Inferences About More Than the Mean of the Population.

In a prototypical sampling distribution, data is normally distributed causing the mean of the sampling distribution to be representative of both the median and the mode of the sampling distribution. However, sampling distributions do not provide information about the how the mean and the mode of the population distribution compare to mu. The mode of the population data is the most probable prediction for a single statistic value of an individual. Thus, when students infer that the population mean behaves like the population mode, they may be invoking the normal distribution imagery. The conclusion that the confidence interval provides a range in which individual values of the population occur most often is sensible and correct only in the case that the population's shape is normal.

The first misconception was identified in the first interviews with Anna and with Ethan as shown in Table 7.

Table 7. *Misconception 1 observations*

	Misconception Observed in First Interview	Misconception Observed in Second Interview
Tiana	No	N/A
Danielle	No	No
Anna	Yes	No
Corinne	No	No
Ethan	Yes	No

Anna and Ethan described that a purpose of a confidence interval is to provide information about individuals in the population and the mode of the population in addition to providing information about the population mean. However, a confidence interval cannot definitively tell us about the behavior of individuals in the population or about the mode of the population. The justifications used by Anna and Ethan indicated that normal distribution imagery was likely activated in their casual nets. The tendency to activate normal distribution imagery may have led them to their overgeneralized conclusions. While activating normal distribution imagery can be necessary and appropriate in some contexts, it can be problematic if its activation excludes the consideration of other types of distribution shapes. Anna and Ethan made inferences that are reasonable assuming that the population is normally distributed, and it was not their tendency to consider any case other than a normal distribution.

1a. Evidence. The following excerpt is from the first interview with Anna. She demonstrated a conception that confidence intervals allow her to make inferences about individual data points in the population. Anna's description of the mean referred to picking a single individual rather than to the average of the population. Again, her inference holds in the case that the population is normally distributed, but fails in the case that the population has a non-normal shape, which she did not consider.

Interviewer: What answer are we going for with a confidence interval?

Anna: Okay, so we are trying to find some mean like where something is most likely to occur or some event is most likely to occur. The mean will give you like averagely what if you, you know, most often if you picked a student what are his sleeping patterns. [] I feel like we use [confidence intervals] so that we can look at one person.

Anna's apparent causal net linked the mean to the value that occurs most often for the individuals in the population since she infers that the mean is "most likely." This could be an appropriate inference if the value of the mean was equal to the value of the mode. The mean and the mode will have approximately the same value if the population data is normally distributed. Anna's inferences appeared to be based on an implicit assumption the population was normally distributed. She did not provide evidence of considering a case with a non-normal population. If the shape of the population distribution is considered then links from the mean to the most likely value are appropriate. A causal net that better approaches an expert conception could link the mean to the mode *if the data is normally distributed* then the mode to the most likely value.

Ethan demonstrated that he thought confidence intervals reflect the mode of the population data. He did this in the context of finding average hand lengths. Ethan's population in this context was not necessarily normally distributed, so the mean, median, and the mode could have had notably different values. Ethan did not draw a population distribution for this context yet made an inference that only holds if the population distribution is normal. He said that "companies that manufacture gloves [could use a confidence interval] to determine the most popular glove size." Ethan's inference would be false if adult hand sizes follow a bimodal distribution like the trend of heights where one peak is the mode for females and a second peak is the mode for males. If the distribution is bimodal then the most popular glove size would not be the mean since the mean would be situated in the valley between the two peaks.

1b. Connections to Classroom Experience. When students discussed μ in ways that would be more representative of the mode of the data without checking for normality, I recognized it as a tendency to imagine normal distributions. For example, Anna referenced an experience in her Introduction to Statistics class in which the instructor had placed a sampling distribution of $\bar{x}$ over a population distribution to demonstrate that both of their means coincide at the same value. Another purpose of the instructor's demonstration could have been to show that the spread of the sampling distribution of $\bar{x}$ is smaller than the spread of the population, though Anna did not state this. When Anna drew an image to mimic her teacher's demonstration I noticed that both distributions were normally distributed. (See Figure 1). An unintended consequence of this demonstration was that Anna visualized μ at the peak of normally distributed data and did not appear to consider that data could be distributed so that μ is not at the peak, or mode, of the distribution. It is reasonable to believe that this image encouraged her to assume properties of μ include properties of the mode because that is true for normal distributions.

Interviewer: Could you tell me more about where you would identify "the answer" if you were looking at a distribution?

Anna: The confidence interval, well, just… I'll draw a picture. So my teacher in [Introduction to Statistics] showed us this (draws a normal distribution with a second, skinnier normal distribution overlapping it shown in Figure

1) so I know I want the mean, like the middle of the data so we know what happens most often.

Interviewer: Is that place where the distribution is the highest the answer that all confidence intervals try to find?

Anna: Yes.

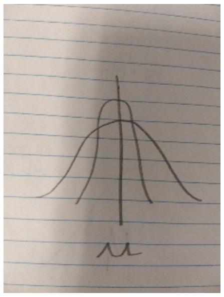


Figure 1. *Anna's drawing of two overlapping normal distributions.*

The conclusions that Anna and Ethan made when they interpreted confidence intervals assumed the case of a normal population, and they did not discuss the possibility of other population shapes during their first interviews. Table 8 is intended to expound on nuances that are necessary to understand for appropriate links to be formed concerning when μ has approximately the same value as other characteristics of the sampling distribution or population distribution. These nuances were possibly not emphasized in these students' experiences with Introductory Statistics in a way that allowed the students to notice their significance. Where the table entry is "Yes" it indicates that the population mean and the table entry's column label can be assumed to be equal or approximately equal given the condition of the associated row. Where the table entry is "No" it indicates that the population mean and the table entry's column label cannot be assumed to be equal or approximately equal given the condition of the associated row.

Table 8. *Comparisons to the population mean given conditions of normality or C.L.T.*

	a. Population median = population mean	b. Population mode = population mean	c. Sampling distribution mean = population mean	d. Sampling distribution median = population mean	e. Sampling distribution mode = population mean	f. A sample distribution's mean = population mean
If the population is Normal	Yes	Yes	Yes	Yes	Yes	No
If the population is not Normal and C.L.T.* applies	No	No	Yes	Yes	Yes	No
If the population is not Normal and C.L.T.* does not apply	No	No	Yes	No	No	No
*C.L.T. stands for Central Limit Theorem						

It is worthwhile to note that in column c, the answer is always "Yes." In other words, the mean of the population will be equal to the mean of the sampling distribution regardless of the shape of the population distribution. It may be that while hoping to stress this phenomenon, Anna's statistics teacher unintentionally taught her to disregard the shape of the distributions. She seemed to reason with only the case of a Normal population in mind without awareness of the cases listed in subsequent rows. I believed that because at one point in the interview Anna made her use of the normal distribution explicit and justified the use of only normal distributions by citing the Central Limit Theorem.

Anna: I don’t think my reasoning is very good but we base a lot of things off of being symmetric. (She draws a normal distribution). With a normal distribution curve it actually turns out a lot of times, like in life with probability, like that's how things end up.

Interviewer: Symmetric?

Anna: Yeah. And there's some variance but if you have a really high chance of getting it like you're going to get that most of the time so it just kind of works out that way and like the less probable it is, the less likely you're going to get it so the data just follows this normal distribution. Like we did in class with the Central Limit Theorem.

Anna assumed the properties of a normal population distribution apply (specifically, that the population mean, median, and mode are equal) if the sample size is at least thirty. She incorrectly applied the Central Limit Theorem to conclude that her population data was approximately normal rather than that the sampling distribution was approximately normal.

1c. Resolution. Anna moved toward a more expert conception about appropriate inferences that can be made from a confidence interval when she chose a context in which she explicitly used skewed population data. She introduced a rainfall context for which she drew a skewed population distribution as shown in Figure 2. In this excerpt Anna’s attention was drawn to the fact that having a skewed population distribution causes the mean and the median to occur at different values.

Anna: Yup. So like going to the left of 25 there’s more data points. Like there’s just more area covered here versus like… Which is interesting! So then why would I choose something higher? More points are over here. But the mean is pulled the way of the skew, right? Cuz these [left side of right skewed distribution] are bigger numbers that are going to affect your average more. [] If our data is skewed we could have more standard deviations with data above the mean than below it. So μ doesn't have 50% of the data above it and 50% of the data below it.

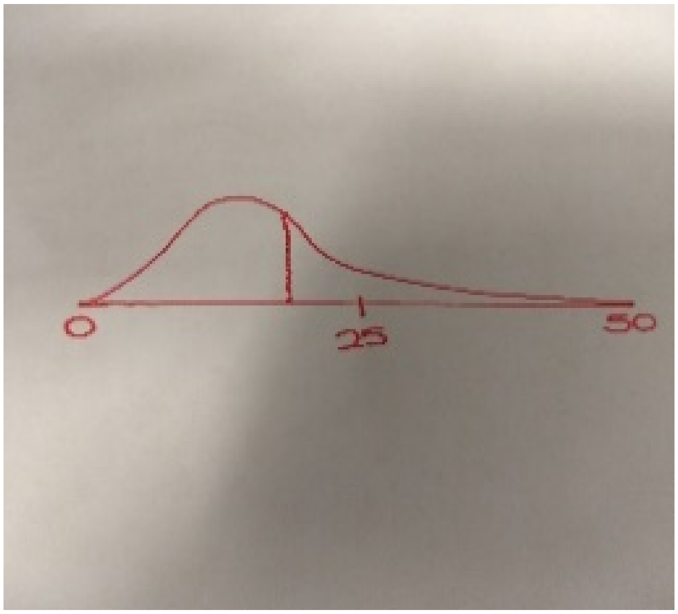


Figure 2. *Anna's skewed population image.*

Another instance in which Anna's conception may have approached a more expert conception was in her second interview with another case of skewed data. I created a data set to not follow a normal distribution so that she could find unique values for the mean, median, and mode of the data set. The data set I gave her contained the values 1, 2, 3, 4, and 40.

Interviewer: How would you calculate the mean?

Anna: Well you would add all these together. That came out really nice. I think it's ten.

Interviewer: Ten is the median or the mean?

Anna: Ten is the mean. (She crosses off 1, 40, 2, and 4). Three is the median. So those are pretty different numbers: Ten and three. And they all occur just as often. Like you have to use your data set. The interval is not centered around the population mean, median, or mode. It will try to approximate the population mean which might match it, but [the confidence interval] is not built to approximate the median or the mode.

This was an example of moving toward an expert conception because she concluded that the confidence interval attempts to capture the population mean, but it is not constructed to capture the population median or mode. For the median and mode, she said one would have to use the original data set. That statement marked a resolution in the given context to her earlier misconception because she realized her inferences about the population median and mode were dependent on the shape of the population.

In summary, at the beginning of the interviews Anna expressed that the purpose of a confidence interval was to tell her about the mode of the population and to make predictions about one person. She calculated the mean, median, and mode of a small, skewed data set to arrive at the conclusion that one must use the original sample data rather than the confidence interval in order to make inferences about the median and the mode of the population.

Misconception 2: The Range Spanned By a Confidence Interval with a 95% Confidence Level is the Range that Contains 95% of the Population Data.

This second misconception is that some students believed that the purpose of a confidence interval is to indicate a range in which the middle CL% of the population data is situated. Oftentimes I have chosen to use 95% as a specific CL% in this paper for ease since 95% is a commonly used confidence level in practice. Further, approximately 2 standard deviations away from the mean of a normal distribution contains 95% of data, and “approximately 2” is a fairly convenient value since it is close to an integer.

This misconception was identified in the first interview with Tiana and the first and second interviews with Corinne as shown in Table 9.

Table 9. *Misconception 2 observations*

	Misconception Observed in First Interview	Misconception Observed in Second Interview
Tiana	Yes	N/A
Danielle	No	No
Anna	No	No
Corinne	Yes	Yes
Ethan	No	No

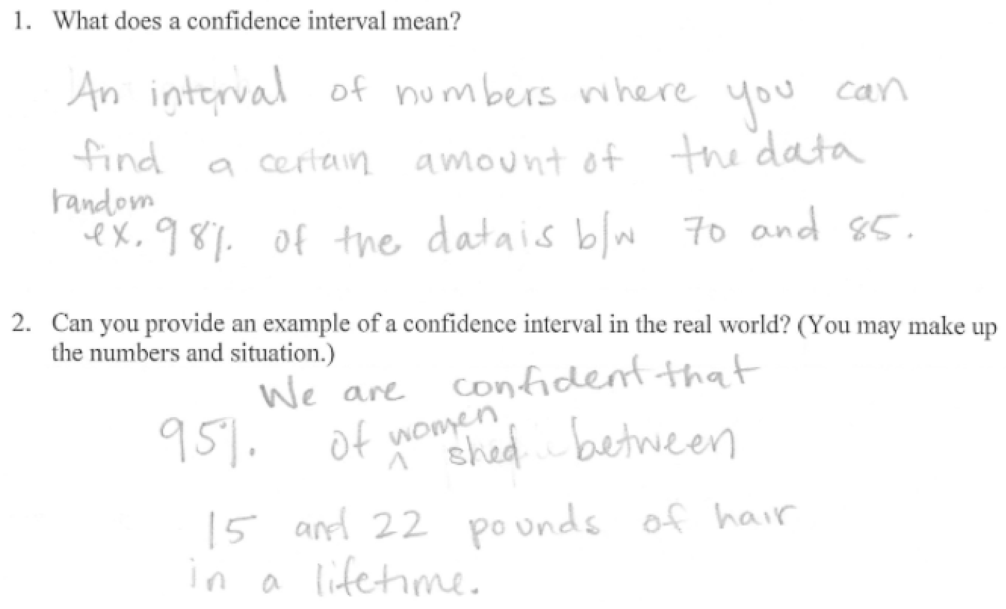
Tiana and Corinne persisted to exhibit this misconception for levels of confidence other than 95% such as 90% and 80%, but 95% was used in the title of this misconception because it is common. Readers may appropriately substitute other examples of confidence levels besides 95% and assume the results found with a 95% level are still applicable.

2a. Evidence. Evidence for this misconception arose when I asked the students to describe the process of calculating a confidence interval. For example, Corinne came up with a quantitative variable of distance traveled by a paper airplane on her pretest response, and she exhibited the misconception when I asked her to explain that scenario. She explained that the process of calculating a confidence interval includes lining up all of the data points in increasing order and then breaking them into percentages so that 2.5% of the data points would be below the lower boundary and 2.5% would be above the upper boundary. This procedure clearly constructed the confidence interval so that it would contain the middle 95% of the sample data.

Interviewer: How would you do this if you were in charge of a study with airplane throws and you had to build that confidence interval yourself?

Corinne: I would have them thrown in the same way to control as many factors as possible. I would record how far they went and organize [the data] from smallest to largest. Find the boundaries so that between all those values is 95% of the data. Then 2.5%. So then you get the 2.5% on either end if you are using 95%.

Tiana also indicated that she had this misconception. She reviewed responses that she had written on her original pretest which both described that confidence intervals contain a specific percent of the total data. Her pretest response is shown in Figure 6. She continued to stand by her pretest responses during the interview. The first pretest prompt asked, "What does a confidence interval mean?" Her response was, "An interval of numbers where you can find a certain amount of the data. A random example is 98% of the data is between 70 and 85. The second pretest prompt asked, "Can you provide an example of a confidence interval in the real world? (You may make up the numbers and situation.)" Her response was, "We are confident that 95% of women shed between 15 and 22 pounds of hair in a lifetime."



1. What does a confidence interval mean?

An interval of numbers where you can find a certain amount of the data

random
ex. 98% of the datais b/w 70 and 85.

2. Can you provide an example of a confidence interval in the real world? (You may make up the numbers and situation.)

We are confident that 95% of women shed between 15 and 22 pounds of hair in a lifetime.

Figure 3. *Tiana's pretest responses.*

After reviewing her pretest response, Tiana was asked to explain her thinking. She did this verbally and drew a picture.

Tiana: So it's this. For example if this were like 98% of all our results or whatever inside the graph we're between 70 and 85 whatever.

Tiana drew the picture shown in Figure 4. The picture she drew was a normal distribution. She marked a position at approximately two standard deviations below the mean and labeled it 70. She then marked a position at approximately two standard deviations above the mean and labeled it 85. Lastly, she wrote "90%" inside the distribution between the marks for 70 and 85. Her original paper had 98 as the percentage and she had also verbally said 98%, so the 90% that was written in this image was actually intended to be 98%, but Tiana wrote 90% by mistake. Tiana then began using the distribution she had just drawn to reason through her explanation, so she started to say 90% like she had written on the distribution.

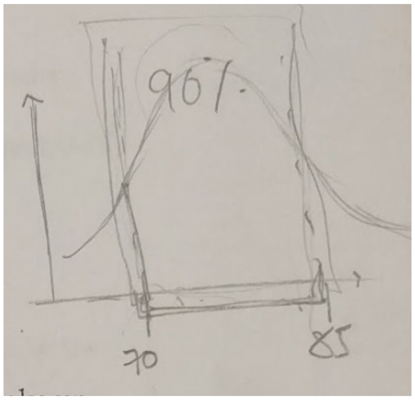


Figure 4. *Tiana's drawing of results between 70 and 85.*

Tiana: So we're saying that 90% of these results lie in here within 90... in the center 90%.

Interviewer: And when you're saying "results"…?

Tiana: From, yeah, results from a sample whether it's one sample or different samples depending on how you do that.

Interviewer: You're showing me the middle 90% of all the data? We're just saying the confidence interval is capturing the middle 90% of the data?

Tiana: It is.

It was unclear whether Tiana meant 90% of the sample data or 90% of the total data, but in either case she was still exhibiting a misconception. Since the boundaries for the middle 95% of the sample data could serve as an approximate for the boundaries of the middle 95% of the population data, she may have meant both cases.

2b. Connections to Classroom Experience. Students think in sensible ways, so when misconceptions are formed it is sometimes the result of images frequently seen in the classroom and applied by the student in inappropriate contexts. For example, the normal distribution that

Tiana drew (see Figure 5) led her to reasonably conclude that 90% of the data fell within the confidence interval.

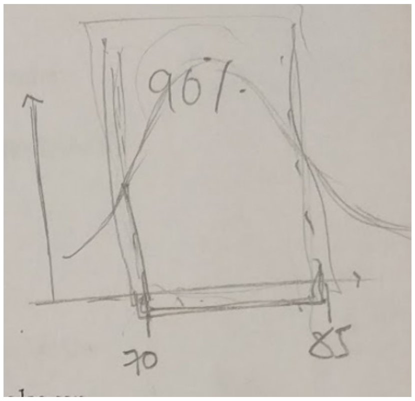


Figure 5. *Tiana's drawing of a 90% confidence interval.*

The image that Tiana drew is a prototypical image used in many textbooks when the topic of confidence intervals is being introduced. The prototypical image shows a normal distribution with equally spaced boundaries on either side of the mean and shades the area between the boundaries (see Figure 6). The prototypical image could have served to reinforce Tiana's misconception that confidence intervals contain a specific percent of data because the image emphasizes that the percent of data inside the interval corresponds to the CL%. However the "data" in the textbook images were the sample means within the sampling distribution, not the individual data points in the sample or the population. Further, the textbook's "interval" is the one centered at the unknown population parameter, not the interval centered at an $\bar{x}$. Tiana's drawing in Figure 5 appeared to be centered at an $\bar{x}$ on the population distribution instead of a sampling distribution which could indicate that she did not understand these critical nuances.

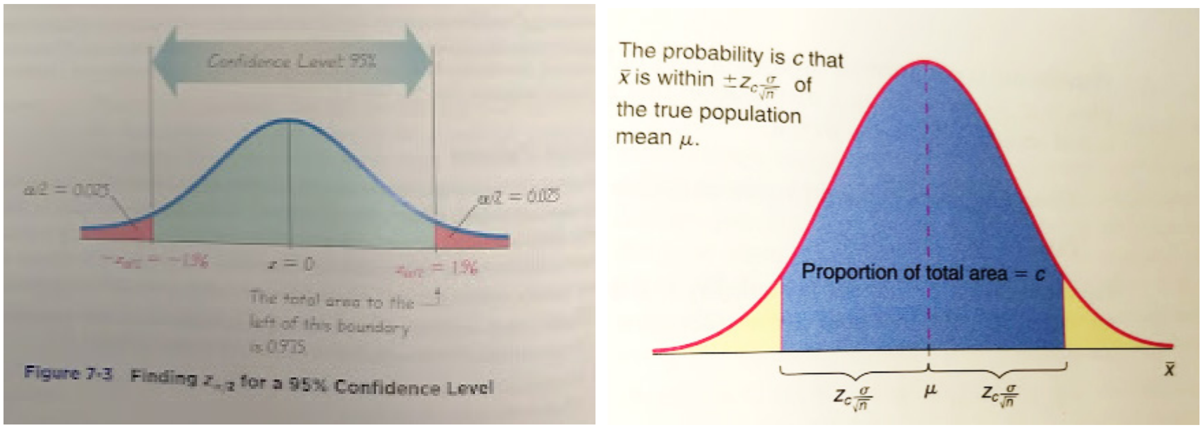
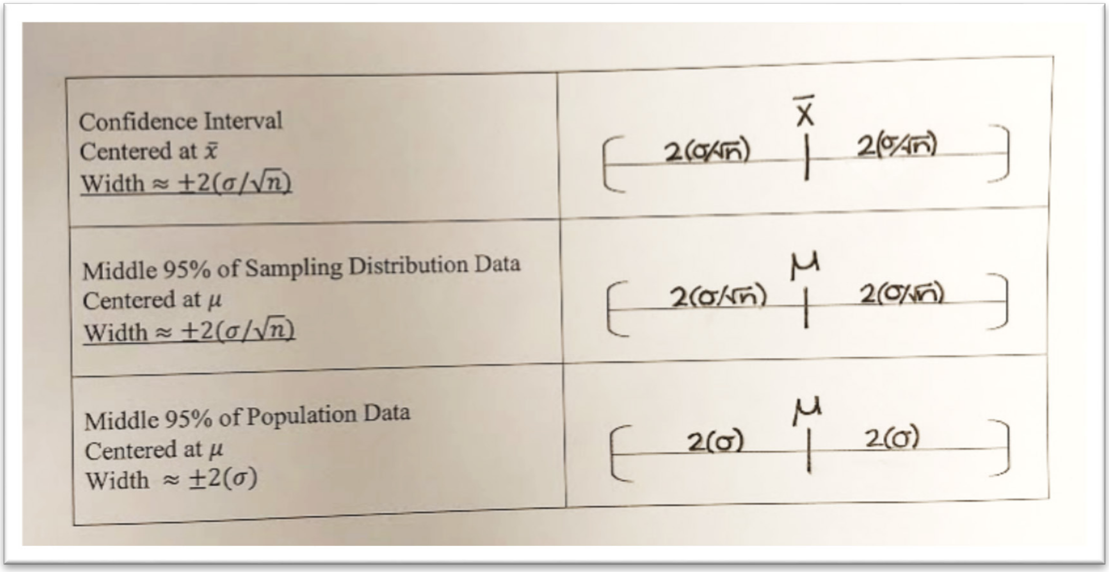

Confidence Level: 95%
$\alpha/2 = 0.025$
$\alpha/2 = 0.025$
$-z_{\alpha/2} = -1.96$
$z = 0$
$z_{\alpha/2} = 1.96$
The total area to the left of this boundary is 0.975
Figure 7-3 Finding $z_{\alpha/2}$ for a 95% Confidence Level
The probability is c that $\bar{x}$ is within $\pm z_c \frac{\sigma}{\sqrt{n}}$ of the true population mean μ.
Proportion of total area = c
$z_c \frac{\sigma}{\sqrt{n}}$
μ
$z_c \frac{\sigma}{\sqrt{n}}$
$\bar{x}$

Figure 6. *Sampling distribution images used in textbooks.*

Despite their misconceptions, Tiana and Corinne discussed the variability of $\bar{x}$ in sound ways. It was not the case that Tiana or Corinne thought that the sampled $\bar{x}$ could not vary from μ since they did not indicate that their confidence intervals were centered right at μ. Regardless of $\bar{x}$'s proximity to μ, both the confidence interval and the middle CL% of the sampling distribution span the same distance. Instead of mistaking the concept of variability, a major contributor to Tiana and Corinne's misconception seems to have been the width of the confidence interval. Specifically, they reasoned about the width of the confidence interval in relation to the range containing the middle CL% of "the data." As indicated at the beginning of this section, depending on whether "the data" is referring to data on the sampling distribution or on the population distribution, the reasoning could be indicative of a misconception or of an expert conception. Figure 7 uses a 95% level of confidence as a specific example intended to highlight that the width of a confidence interval is the same as the width of the range containing the middle CL% of the sampling distribution's data. Although Figure 10 shows each of the respective ranges covering the same distance, the distance spanned by the middle 95% of the population data would be greater than the other two ranges because 2σ is greater than $2(\sigma/\sqrt{n})$. This discrepancy may be obscured verbally as “two standard deviations” since both σ and $\sigma/\sqrt{n}$ measure standard deviations.



Confidence Interval Centered at $\bar{x}$ Width $\approx \pm 2(\sigma/\sqrt{n})$	$\bar{x}$ $[\ 2(\sigma/\sqrt{n})\ \vert\ 2(\sigma/\sqrt{n})\ ]$
Middle 95% of Sampling Distribution Data Centered at μ Width $\approx \pm 2(\sigma/\sqrt{n})$	μ $[\ 2(\sigma/\sqrt{n})\ \vert\ 2(\sigma/\sqrt{n})\ ]$
Middle 95% of Population Data Centered at μ Width $\approx \pm 2(\sigma)$	μ $[\ 2(\sigma)\ \vert\ 2(\sigma)\ ]$

Figure 7. *Comparisons of the width of various ranges*

An important element of the prototypical image (as in Figure 6) is that the boundaries on the shaded region are constructed using the same margin of error found in the confidence interval. The width in the image is the same width of the confidence interval. Tiana and Corinne may have read out the prototypical image of confidence intervals on the samplings distribution of $\bar{x}$ and inferred that confidence intervals contain CL% of the data. In doing so, they may have believed that the distribution in the prototypical image is either the population or the sample distribution (not the sampling distribution).

2c. Resolutions. Corinne persisted with this misconception into the second interview, but revised her conception about how the confidence interval is calculated when I pressed her to describe how her method worked with skewed data. At first her method was based on arranging the data from smallest to greatest, identifying the percent of the data that would be excluded from the interval (such as 5% when using a 95% level of confidence) then splitting that percent in half (5% divided into 2.5%) and setting the boundaries for a confidence interval just after the first 2.5% of the data and just before the final 2.5% of the data. When she revised her method it was because she determined she wanted to identify the center of the interval first. She initially used the mean as the center, then she decided the median was a better fit for data that was not symmetric. She made sure that the difference between the value used for the lower bound of the interval and the median was the same difference as between the value used for the upper bound and the median. Having equal margins of error on either side of her estimate took precedence over ensuring that the confidence interval excluded the same percent of the data on the upper and lower ends in her revised method.

Interviewer: What if your data wasn't symmetric?

Corinne: You would center it around the median [instead of the mean].

Interviewer: How do you set the boundaries?

Corinne: Whatever distance this one has to go to capture 80% of the data. And whatever that distance is, this other side's boundary would just mimic it kind of like an absolute value sort of thing.

Since Corinne's reasoning shifted to favoring equal margins of error over equal percentages of the data excluded on either side of $\overline{x}$, she approached a more expert conception. She did this by making inferences about the center of confidence intervals.

Misconception 3: Calculations for Confidence Intervals are Not Based on Actual Sampling Distributions.

Recall from chapter 2 that a *sampling distribution* is a distribution of sample statistics, such as sample means. Generally, sampling distributions are defined to have an infinite number of samples. This is because in practice, populations are changing over time. For example, if the population of interest is Americans with hypertension, the population technically includes a finite number of people right now. However, the population is going to be different five minutes from now, and because the population is continuously dynamic it is considered to be infinite. Further, the process of calculating sample means may be considered to have no end because resources of time and money will be exhausted before the repeated sampling process can be "finished." Thus, the associated sampling distribution is considered to contain an infinite number of samples.

In this paper I make a distinction between the sampling distribution as normatively understood by statisticians, and a different approach to sampling distributions evidenced by the student misconceptions in my study. I will use the term *actual sampling distribution (ASD)* to mean the object that is the encapsulation of the process of taking all possible samples and plotting a sample statistic from each sample. That is, the ASD is what statisticians refer to as *the* sampling distribution. I contrast this with the students' conception of a *sampling distribution process (SDP)* which is the process of repeatedly adding more sample statistics to a distribution of a potentially infinite number of sample means without encapsulating the ASD as an object.

Each of the students interviewed had an inaccurate notion that statisticians apply a version of the sampling distribution that approaches the ASD, but does not contain every combination of a sample of size n from the population. In many instances, they suggested a specific, finite number of sample means in the sampling distribution. A distribution of sample means that contains many means, but not all means will be called a *Multi-Sample Distribution (MSD)*. The MSD represents the truncation of the limiting process that would lead to an ASD. For every sample mean that could be added to the SDP, there exists an MSD. A key idea for the students was that the MSD was required to be practically doable in real-world situations, meaning it could not contain more samples than a researcher would realistically gather.

AN MSD is to the ASD as a sample is to the population. Both the ASD and the population contain every possible combination or individual. Neither the MSD nor the sample will include all the combinations or individuals that could possibly be sampled, just a subset. To better illustrate the ASD, the SDP, and the MSD, I will give a metaphor utilizing limits. Consider $\lim_{n \to \infty} \frac{n}{n+1}$. The metaphor associated with the ASD is final result from the limiting process which is the numeric value 1. The metaphor for the SDP is the unbounded process of incrementing n to generate successive terms that approach the limit (namely $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5} \ldots$). The MSD is analogous to any single term in the list of sequence values (except the first and last terms) such as $\frac{3}{4}$ or even $\frac{n}{n+1}$. The term used as the MSD depends on the number of samples that the student considers to be a reasonable number of samples a statistician could take in real-life. This is subjective similar to how the term considered to be a “good enough” approximation of 1 can vary for people.

The third misconception was identified in the first interviews with Tiana, Anna, Corinne, and Ethan as well as the second interviews with Danielle, Anna, Corinne, and Ethan as described in Table 10. Hence, all of the students in this study, at one point or another, showed disbelief in the ability to use the ASD.

Table 10. *Misconception 3 observations*

	Misconception Observed in First Interview	Misconception Observed in Second Interview
Tiana	Yes	N/A
Danielle	No	Yes
Anna	Yes	Yes
Corinne	Yes	Yes
Ethan	Yes	Yes

I considered the students' lack of an encapsulation of the ASD to be a misconception. This is a misconception because they did not see the object associated with a sampling distribution. Some researchers have described a related misconceptions about limits such as the misconception that a limit is an unending process with no associated limit value (Davis & Vinner, 1986; Jones, 2015). Since the sampling distribution involves a limit, it may be viewed as consisting of separate 'process' and 'object' components just as the limit concept has been. Researchers note a key difficulty exists in helping students see that there is an 'object' associated with the process of a limit (Davis & Vinner, 1986; Jones, 2015).

3a. Evidence. I documented instances in which students used an MSD as evidence of the misconception that calculations for confidence intervals are not based on ASDs. Students used the MSD when they did not see the utility of the ASD. The utility of an ASD was often obscured by the belief that since ASDs are not created in practice by actually producing all possible samples (which would be infinitely many samples), they cannot be used. Although logical, that is a misconception because the ASD is applied theoretically when using confidence interval estimation procedures.

Corinne showed evidence of reasoning with the MSD because she believed that a sampling distribution does not have all possible samples. She described the SDP where each iteration of the MSD had a mean that approached the mean of the population.

Corinne: If you took a bunch of samples and you found their means, you would get a sampling distribution.

Interviewer: How many sample means?

Corinne: More than 30.

Interviewer: Is there a different sampling distribution that could be created for every number of samples?

Corinne: Yes, but after you get 30 they are basically the same thing and will all be approximations of a normal distribution. The mean of the means approximates the theoretical mean. If we're going to get the real mean you literally have to sample everyone in the population. The more samples, the more normal. In the real world we never get to work with the distribution where μ and $\mu_{\bar{x}}$ are equal. We just get closer and closer.

Corinne's plausible causal net linked a bunch of sample means (but not all sample means) to the formation of a sampling distribution. This is incorrect if the view of sampling distributions is limited to the ASD, but her definition fits well with the definition of MSD. Corinne also indicated that if there are more than 30 samples, she can infer that the shape of the sampling distribution is approximately normal. In this, Corinne seems to be drawing on the productive idea that a *single sample* of size 30 or more is often considered to be large enough for certain statistical techniques, such as using the *z*-distribution with $\sigma \approx s$ instead of the *t*-distribution. However, she has applied this oft-cited rule to the wrong context in claiming that having at least

30 samples is sufficient to assume one has a complete sampling distribution. Further, her inference then led her to conclude that the mean of her sample means would approximate the mean of the population. The mean of an MSD would approximate the mean of the population. However, Corinne has a misconception because in normative understandings, “the sampling distribution” is the ASD, but in Corinne’s understanding, “the sampling distribution” is the MSD of at least 30 samples. The mean of the ASD does not approximate the mean of the population since the mean of the ASD is exactly equal to the population mean. Corinne indicated that in the real world we cannot sample every sample combination in the entire population which implied that in practice we can only approximate the ASD. Table 11 summarizes portions of Corinne’s causal net in this context.

Table 11. *Likely causal nets demonstrated by Corinne*

A bunch of sample means	→	A sampling distribution (MSD)

More than 30 sample means	→	Sampling distribution (MSD) approximates Normal distribution	→	Mean of sampling distribution (MSD) approximates population mean

Real world	→	Cannot create the sampling distribution centered exactly at μ (ASD)	→	Can only get closer and closer (SDP)

Like Corinne, Tiana also described an MSD because she explained that the average of all the samples on a sampling distribution is close to μ, but does not equal μ.

Tiana: When we do that distribution it's like a plot of the $\bar{x}$. When we take more of those sample sizes we can get closer to the true mean.

Interviewer: How do the $\bar{x}$ on that distribution compare to the true mean?

Tiana: From all those $\bar{x}$ wouldn't [μ] just be the average? $\mu_{\bar{x}}$?

Interviewer: So the average of all the averages is μ?

Tiana: Kind of. I'd say μ would just be close.

The description of a SDP fits well with Tiana's description since she described a plot of the sample means with a mean that gets closer to the true mean as more sample means are included in the distribution. Therefore, the SDP is likely linked in a causal net for Tiana as shown in Table 12.

Table 12. *Likely causal nets demonstrated by Tiana*

The sampling distribution process (SDP)	→	A plot of the $\bar{x}$'s whose center gets closer to μ as more samples are taken	→	The center approximates μ, but will not be equal to μ

It is true that with more sample means included in any given MSD, the mean of the distribution is likely to be closer to μ. However, it is also true that successive sample means could make the average of the MSD with more than x samples less close to μ than the average of the MSD with x samples. Hence, increasing the number of samples in the MSD will not always coincide with a better approximation of μ or with more normally distributed data than the previous MSD with one less $\bar{x}$. The mean of a MSD is not guaranteed to be equal to μ until the entire process of adding sample means is complete and the distribution becomes the ASD. Since

Tiana thought of a sampling distribution as a SDP rather than as an ASD, she was technically correct that the mean of an MSD will approximate μ but not equal it.

Danielle also demonstrated that she did not think of the sampling distribution as having every possible sample, just many samples. However, unlike Tiana, Danielle thought that the MSD would be centered at μ while Tiana thought the center would just be close to μ.

Danielle: When you take samples if you have a large enough number of samples from skewed population then the distribution of those means in a sampling distribution will be approximately normal and will be centered at the population mean.

I believed she was reasoning with the MSD because she reasoned with an arbitrary number of samples without referencing the iterative pattern of taking more samples or getting "closer and closer" to anything (which would be the SDP) and without the distribution including every possible sample combination (which would be the ASD). Her conclusion that the sampling distribution is centered at μ, as shown in Table 13, is problematic given that she was likely thinking of the MSD rather than of the ASD.

Table 13. *Likely causal nets demonstrated by Danielle*

Given a large number of samples	→	The sampling distribution (MSD) is approximately normal	→	The sampling distribution (MSD) is centered at exactly mu

Ethan described an MSD when I asked him to imagine he was in charge of a fictitious study. I asked him to plan a study and to be as efficient with his resources as he could. His understanding was that if only one sample was taken then he would only be able to use a distribution with one sample. He decided that a distribution with one $\bar{x}$ can still qualify as a sampling distribution with one point.

Interviewer: Could you get away with a sample less than 600?

Ethan: How confident am I trying to be here?

Interviewer: Technically we could just a sample of one.

Ethan: That's ridiculous. But what if we only do 30.

Interviewer: Why is 30 a magical number to you?

Ethan: The Central Limit Theorem wants 30 for the sampling distribution to be normally distributed.

Interviewer: What's a sampling distribution?

Ethan: The distribution of the means of your samples.

Interviewer: How many?

Ethan: Are you talking about in my thing or just in general?

Interviewer: A sampling distribution is just defined one way isn't it? What is the Central Limit Theorem talking about? What distribution?

Ethan: Of the means of how many samples you take.

Interviewer: But what if you only take one?

Ethan: If you only take one sample then... I'm lost.

Interviewer: You decided the sample size of 30 was special to you because of the Central Limit Theorem and because of Central Limit Theorem the sampling distribution will be normal.

Ethan: Yes.

Interviewer: What on earth is a sampling distribution?

Ethan: That's the distribution of the means of all your samples.

Interviewer: It is the distribution of the means of all your samples. Okay. All the samples that you take or all the samples that you could take?

Ethan: That you take.

Interviewer: K, so if we only do this once, with just one sample of size 30 like you suggested, do we have a sampling distribution?

Ethan: You've only got one mean.

Interviewer: So the sampling distribution doesn't exist?

Ethan: Well, it's a distribution of one and you can call that a distribution.

Ethan's remarks indicated that he thought of the sampling distribution as the MSD because he defined it as the distribution of means of samples you take. This is shown Table 14. Also related to his concept projection of sampling distributions were that a sample of size 30 evoked the Central Limit Theorem and therefore created a normally distributed MSD and that a sample of size one is ridiculous.

Table 14. *Likely causal nets demonstrated by Ethan*

Sample size of 30	→	Central limit theorem	→	Sampling distribution (MSD) normally distributed

A sample of size n=1	→	Ridiculous

The sample distribution (MSD)	→	Distribution of means of samples you take

3b. Connections to Classroom Experience. A specific classroom activity was described by Tiana that encouraged her to think about the sampling distribution as an MSD instead of an ASD. The activity was for the students to compute a sample mean for one sample and then all of the students plotted their sample means together on the same distribution. I have used an activity like this myself, and the goal of the activity was to show students where the sampling distribution of $\bar{x}$ comes from. What we literally build, however, is an MSD. When Tiana saw the MSD created in class, her readout may have been that she was viewing the actual sampling distribution. If so, the plotted distribution of multiple sample means was her center of focus. Although Tiana's instructor may likely have explained that the sampling distribution was not the MSD image she was seeing, but rather an ASD, this nuance could have been easy for Tiana to miss. After all, Piaget described the tendency of students to notice only one salient aspect of a situation at a time to the exclusion of other potentially relevant aspects (Piaget, 1952). A second center of focus that Tiana recalled from her classroom experience was the center of the MSD. Tiana's readout was that $\bar{x} \approx \mu$. A second possible readout for that image would be that $\bar{x} = \mu$. This second readout is conducive to explaining Danielle's reasoning in the previous subsection about the mean of the means equaling μ on an MSD. Yet, in many ways, at the level of an MSD, it would be more accurate to state the mean of the means as $\bar{x}_{\bar{x}}$ because it only contains a *sample* of means and not *all possible* means.

Corinne's interview provided an additional reference to classroom experience that impacted how she thought of the MSD. She discarded knowing the ASD in a manner akin to discarding knowing σ. She did not argue that the sampling distribution does not exist (or that σ does not exist), but she said that in general we cannot get an ASD so it is not usable.

Corinne: $\mu_{\bar{x}}$ is the mean of the means you sampled. You use it to calculate the confidence interval.

Interviewer: How does $\mu_{\bar{x}}$ compare to μ?

Corinne: It should approximate it based on the Central Limit Theorem. If you increase the number of samples then eventually $\mu_{\bar{x}}$ is equal to μ. In the real world we never get to work with the distribution where μ and $\mu_{\bar{x}}$ are equal. We just get closer and closer. It's like we can't get σ and so you get closer and closer to it with s. You could only get the means to be equal in something like manufacturing where you have data on every item or when you have a small population but in that case it would be pointless because you could just do a census and know the population parameters.

In the real world we approximate μ, so Corinne was correct. However, she was not seeing that in the real world we rely on the ASD in theory. Statistics are applied to populations for which a census is not possible, so the purpose of statistics is to estimate parameters. If a situation justifies the need for finding statistics (as opposed to finding parameters), then the population is too large for a census. If a population is too large for a census then it will be far too large for an ASD of sample means to be constructed. This is because every individual in the population is surveyed once by a census, but every individual in the population is theoretically sampled numerous times in a sampling distribution. Why would we assume we have an ASD when μ is unknown? That is even more absurd than assuming we know σ when μ is unknown. It is not reality. Nobody constructs the ASD for a given population else why would they not choose to take a census instead? A parallel between the ASD and the population standard deviation as identified by Corinne is shown in Table 15.

Table 15. *Parallel causal nets for σ and ASD*

σ	→	Calculable if all of the data is available	→	Approximated with s when population is too large

ASD	→	Calculable if all of the data is available	→	Approximated with MSD when population is too large

3c. Resolutions A resolution was considered to be reached when students moved toward an expert conception that known properties of the ASD can be applied without constructing the ASD. At first, the students interviewed seemed to see no utility for the ASD because it is only theoretical. In some cases the students knew they wanted "at least thirty" of something so that the MSD would be approximately normal, but they could not articulate why normality was necessary. The students did not know what problem the ASD "exists" to solve. I say exists in quotations since the ASD only exists in theory.

Linking to an expert conception of variability led Corinne to realize that she knew properties of the ASD without needing to build one. In the context of the second interview, Corinne resolved her misconception about needing an MSD to estimate the sampling distribution.

Interviewer: What happens if someone only picks one sample? And let's make this the smallest possible $\bar{x}$ and this the largest possible $\bar{x}$ (I labeled an upper and lower limit on the x-axis below the all possible $\bar{x}$ distribution).

Corinne: Without even knowing anything about this [the specific values on the sampling distribution] most of them are going to be in the middle 95%. So chances are that this one single data point, if we have just one $\bar{x}$, it's here somewhere (gestures to the middle of the distribution).

Interviewer: Okay. Is that helpful?

Corinne: I mean, if that is the case and we know that $\mu_{\bar{x}}$ is equal to μ or approximates it really closely so the center of this [distribution with all possible $\bar{x}$] that this $\bar{x}$ [from a single, random sample] is going to be kind of in this range [by $\mu_{\bar{x}}$] then that means that the confidence interval we're building around [$\bar{x}$] will, there's a better chance it will contain $\mu_{\bar{x}}$ and therefore mu.

Interviewer: Is μ the same as $\mu_{\bar{x}}$?

Corinne: I think at this point they are the same.

Interviewer: Why?

Corinne: Because at this point if we have taken every possible sample, and take their means, and we are finding the mean of all those means that is mathematically the same as finding the mean of all of those at once which is finding mu. So $\mu_{\bar{x}}$, that's μ of the population. So ya, they should be the same. So if this [$\bar{x}$] is going to be in this range around μ or $\mu_{\bar{x}}$ then $\bar{x}$'s confidence interval is more than likely going to contain μ. So we can just use one of these! You can use one $\bar{x}$ to approximate, to make a confidence interval that probably contains μ.

Interviewer: So you are saying we can use just one sample?

Corinne: But you’re basing it off of information about all possible samples. I’m basing it off of a theoretical understanding of what this one sample looks like most likely compared to all possible samples.

At this stage she was developing an expert conception because she realized that she could use properties of the ASD to compute a confidence interval without having to construct all the possible samples and without having to resort to using an MSD as an approximation like she had done at the beginning of the interview.

Misconception 4: Statisticians Must Take Multiple Samples in Order to Compute a Reliable Confidence Interval.

The fourth misconception is that multiple samples need to be taken in order to create a confidence interval. It appeared that this misconception was a product of misconception 3 because students felt that it was required, in practice, to literally take multiple samples to build an MSD for practical use. Interestingly enough, these students were accustomed to being given information about one sample and calculating confidence intervals with only one sample on their homework and other assignments, yet they still had a misconception that in real life statisticians need to repeat the process of taking samples.

This misconception was identified in the first interviews with Tiana, Danielle, Anna, Corinne, and Ethan as well as the second interviews with Danielle, Anna, Corinne, and Ethan. In other words, this misconception was evident in every interview I conducted as shown in Table 16.

Table 16. *Misconception 4 observations*

	Misconception Observed in First Interview	Misconception Observed in Second Interview
Tiana	Yes	N/A
Danielle	Yes	Yes
Anna	Yes	Yes
Corinne	Yes	Yes
Ethan	Yes	Yes

Some students had different ideas about how multiple samples were used to create confidence intervals. For example, Danielle believed that each sample would have its own confidence interval and that statisticians look to see where the confidence intervals are clustering to make more informed decisions about the mean. Ethan believed that each sample was averaged and that the average of the averages (which he called $\mu_{\bar{x}}$) is the mean that is used to construct one confidence interval based on multiple samples.

4a. Evidence. I looked for instances in which students talked about having more than one sample as evidence of this fourth misconception. For example, In The Teaching and Learning of Statistics and Probability class Danielle asked the instructor “which s” she should use for the confidence interval formula involving $\bar{x}$ plus or minus t* times s over √n. This is evidence that she was considering taking multiple samples because otherwise she would not have had multiple s (sample standard deviations) to choose from. In an interview Danielle described the process that she believed statisticians do in real life to construct confidence intervals and it involved taking multiple samples.

Danielle: You take a large number of samples, like say 1,000, to get a sampling distribution of $\bar{x}$.

Ethan, Corinne, and Tiana likewise claimed that they would take a large number of samples if they were in charge of a real study. Ethan expressed that his choice of the number of samples was dependent on the size of the population.

Ethan: Given [a university's] students as the population, I would take 20 samples of size 50. Given the world population I would take at least 200 samples of size 50 just to be realistic.

Corinne: [Given American adults as the population], I would take 100 samples of size 30.

Tiana: [Given a university's students as the population], I want to take 3 samples of size 100.

4b. Connections to Classroom Experience. As previously mentioned, Tiana described an activity she did in class that I have also used with students to build a sampling distribution of $\bar{x}$ which may negatively influence a student's ability to encapsulate the ASD as an object. However, this activity, as typically done, really builds an MSD since the class members construct their own $\bar{x}$ and plot them on the same graph. As discussed before, this results in only a small, finite number of sample means, instead of all possible sample means required for an ASD. I often followed this demonstration of an MSD by telling my students about how the mean of the sampling distribution is the mean of the population. Tiana's interview made me aware of how some students could be getting the message that we need to take multiple samples. They may confound the MSD with the sampling distribution. This should come as no surprise following a demonstration in which I attributed the properties of the sampling distribution to the physical MSD that we had visually displayed in class. If I had stated verbally that the properties of the sampling distribution would belong to the distribution that we had created if the distribution contained all possible samples, it may not have solved the confusion. There could still be students like Ethan who thought that “all possible samples” could mean all the samples that it was possible for us to have actually taken. Hence, misconception 3 could evolve with students sensibly believing that the MSD is the same thing as the sampling distribution. This could cause misconception 4 as well. After all, as educators, we often teach that the sampling distribution is centered exactly at μ and students might see that this equivalence of means came from taking multiple samples. This could influence why Tiana preferred to take three samples of size 100 instead of one sample of size 300.

Some students believe that we should take multiple samples because it will improve our estimates. Specifically, since the MSD approaches the ASD as more samples are taken, then the mean of the sampled means approaches the mean of the population. This can come from a correct understanding of the Law of Large Numbers. The more data collected, the more reflective the sample statistics will be of the population parameters. Anna’s justification for using multiple samples was consistent with an intuitive understanding of the Law of Large Numbers.

Anna: The more you check, the more accurate you're going to be, so then I think that can also have an effect on the confidence interval.

Interviewer: So with more people, do you mean more people in the sample itself or more samples?

Anna: Both. I just feel like in statistics, the higher the numbers in like gathering data the better off you are.

Another rule that students learn in statistics besides the Law of Large numbers that could lead them to think multiple samples are necessary is the Central Limit Theorem. Students wanted to apply the Central Limit Theorem in order to be able to assume a given MSD was normally distributed. The Central Limit Theorem states that if the sample size is greater than 30 (or another number that is considered to be large enough) then the sampling distribution will be approximately normal. Danielle's justification for needing multiple samples was based on the Central Limit Theorem. She correctly reasoned that the distribution that becomes normal must be a distribution containing multiple samples, but her knowledge was incomplete because she had not progressed to realizing that the ASD and the Central Limit Theorem can be used in theory even if just one sample is taken.

Danielle: We need more samples taken because increasing n in a single sample from skewed population data continues to be skewed. Each individual sample may not be normal, but as you plot those different [means of] samples of that sample size it will give a roughly normal distribution where the mean of each of those samples is clustered around the population's [mean].

Corinne, similar to Danielle, explained that multiple samples are necessary to get a normal distribution and she explicitly cited the Central Limit Theorem. The distribution that Corinne and Danielle each appeared to be referring to becoming normal was an MSD. I inferred this from drawings they made of a bell curve with a specific number of $\bar{x}$'s labeled on the interior of the bell curve related to the number of samples they said they would use if they were in charge of a study.

Corinne: If you took a bunch of samples and you found their means, you would get a sampling distribution.

Interviewer: How many means are on that distribution?

Corinne: It has to be more than 30 using the Central Limit Theorem. If you have a large enough quantity of samples it doesn't matter what n is. The size of n will change the standard deviation size, and the number of samples guarantees that you can use this formula [to compute a confidence interval].

Therefore, Corinne demonstrated that students may mistake the Central Limit Theorem as necessitating a large number of samples rather than necessitating a large sample.

Anna's interview gave me insight into a definition that, like the Law of Large Numbers and the Central Limit Theorem, could encourage students to believe that taking multiple samples is necessary for calculating a confidence interval. This definition is also something students are learning about when they are introduced to confidence intervals, so the rules may be liable to be conflated. The definition is for confidence level.

Interviewer: What is confidence level?

Anna: If we were to create a bunch of confidence intervals over a bunch of samples, our percent of confidence is the percent of intervals that have our population mean lie in that interval. The idea of a confidence level rests off of doing this procedure many times.

Interviewer: The procedure of what?

Anna: Taking a bunch of different samples!

The notion of taking multiple samples can appear when defining level of confidence and an inappropriate link may be formed between confidence intervals and a multiple samples requirement. This is because confidence levels are based on the percent of possible sample means in the sampling distribution that will be close enough to μ to capture μ within their associated confidence intervals. Thus, a confidence level is an indication of how successful the method of calculating confidence intervals would be at capturing the population mean. If we took many samples, we would expect that the number of associated confidence intervals containing μ would be the percentage given as the level of confidence.

Thus, the Law of Large Numbers, the Central Limit Theorem, and the definition of confidence level may all have the unintended consequence of leading students to believe that multiple samples are necessary in order to report reliable confidence intervals.

4c. Resolutions

Danielle, Ethan, and Corinne all concluded their second interviews with some indication that they had changed their minds and instead believed that taking only one sample would suffice. Danielle's change of mind occurred as she plotted a population distribution and a sampling distribution on the same axis. Ethan and Corinne's resolutions both came when they were considering which $\bar{x}$ from the multiple samples would be the $\bar{x}$ they should use to construct the confidence interval.

Danielle stated early in the second interview that unless data is normally distributed then probabilities associated with the Empirical Rule (also known as the three-sigma rule or the 68-95-99.7 rule) and her tables with z* and t* values could not apply. For this reason she believed that it is only appropriate to calculate confidence intervals if the condition of normality is met. This is true. However, to meet the condition of normality she wanted to take multiple samples. Danielle said multiple samples must be taken because "increasing n in a single sample from a skewed population continues to be skewed." Danielle began to resolve her multiple sample misconception when she stated that any sample mean will land somewhere on the sampling distribution. She compared the spread of the sampling distribution to the spread of the population distribution and noted that all $\bar{x}$ values fall on the sampling distribution somewhere, and the sampling distribution's values are clustered about μ. She started to reason about the information that could be assumed from a single sample rather than from building an MSD, and this may have allowed her to think about the theoretical properties of the ASD that will exist even if the statisticians do not build the ASD or estimate it with an MSD. She described the relationships between $\bar{x}$, the sampling distribution, and the population most clearly when she drew them on the same axis. This description likely led her to conclude that only one sample is necessary to compute a confidence interval.

Danielle: Any kind of sample you take is going to fall on the sampling distribution somewhere. It is possible to get one that is farther away from the population mean. If you are looking at the sampling distribution and your population, if you were to graph them on the same axis it would look something like this [Figure 8].

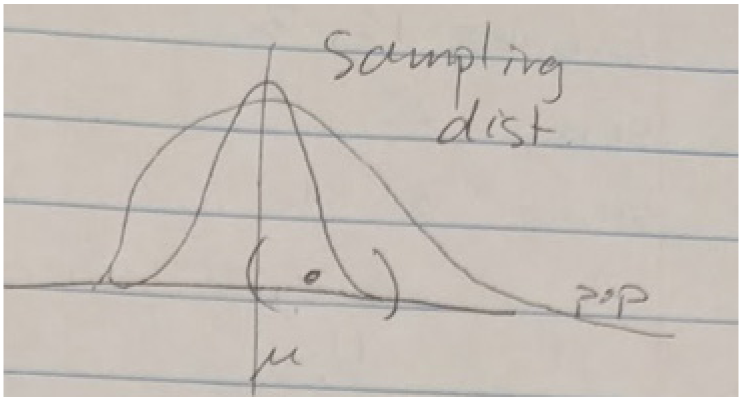


Figure 8. *Danielle's drawing of overlapping distributions.*

Danielle: If you were to take a sample, just one sample, then it will fall somewhere along here in this range that is close to the population mean [on the sampling distribution whose spread is smaller than the population spread] so if you use just one, like when we are making a confidence interval we are just saying that the true mean can be inside that range of the interval.

So if I pick a sample mean and create a confidence interval then the population mean is supposed to fall within there with your level of confidence that you used to create the interval. So you only need to take one.

Ethan and Corinne's resolutions were reached when considering which $\bar{x}$ and s to use to compute a confidence interval after having taken multiple samples. They conceptualized that there would be a mean of the many sample means which they each called $\mu_{\bar{x}}$. When deciding which values to plug into a formula for computing confidence intervals, Ethan considered the mean of many samples as the value for $\bar{x}$ and Corinne considered the mean of many samples as the value of $\bar{x}$ as well as the standard deviation of many samples as the value of s. As a result, both Ethan and Corinne realized that they were averaging their samples' statistics into single values that were representative of all the samples and asked themselves whether they could take one giant sample as opposed to averaging many smaller sized samples.

Corinne's resolution is included in the third misconception's resolution section since she concluded to only take one sample and she moved from reasoning about an MSD to an ASD. Ethan's resolution is included in this section because he resolved to take only one sample, but he continued to think about the MSD as an approximation of the ASD. He seemed to begin reasoning about using the average of all the sample means that were "realistically" collectable when he indicated the formula that he could use to compute a confidence interval.

Ethan: We could get an average of the averages.

Interviewer: What should we call that?

Ethan: $\mu_{\bar{\bar{x}}}$.

Interviewer: Is that the same as the population μ?

Ethan: No.

Interviewer: K. What formula are you going to use to give a confidence interval with that data?

Ethan: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$.

Interviewer: What would $\bar{x}$ be?

Ethan: Weird. Does every $\bar{x}$ have its own confidence interval? I think the formula uses $\mu_{\bar{x}}$. Then take an s which is a standard deviation of one of these samples.

Interviewer: Does it matter which s you use?

Ethan: Not sure. I guess because I haven't heard it so maybe it doesn't matter but to be safe I'd take one from an $\bar{x}$ near $\mu_{\bar{x}}$.

Next in the interview Ethan brought up the Central Limit Theorem. He wanted the sample size to be at least 30 so he would have a normal distribution. He said that a single sample does not become normal since it will follow the shape of the population distribution. I pressed him to tell me what is normal, and he arrived at the conclusion that the sampling distribution (in the sense of an MSD) would be normal and that you could just imagine taking more samples, but only one sample is necessary for constructing a confidence interval and he could collapse the data from multiple samples into one giant sample statistic.

Interviewer: What's normal then? How does n equaling 30 help us?

Ethan: Because if you were to take a lot of samples of n equals 30 then those $\bar{x}$ values would be normally distributed. So 30 is okay because IF you were to keep getting samples then it would lead to a normally distributed

> sampling distribution. It would be really efficient we can just do [a sample of 30] once and imagine that the other samples filled it in.

Expert Conception 1: Variability Exists in the Values of Sample Statistics.

In this study, variation was used appropriately and repeatedly throughout all of the pre-service teachers' verbal justifications of the construction of confidence intervals. I identified examples of this theme any time a student expressed a belief that a value could vary. For example, Anna said, "Since you are talking about bunch of $\bar{x}$ you are going to have a smaller standard deviation because within a sample you can have your numbers be all over the place but when you are averaging all that out you are going to . . . shave[] off the extreme ones and you are not going to get a sample [mean] that is way extreme."

Ultimately, misunderstandings about confidence intervals did not seem to stem from being unable to reason about variation. This was an interesting finding because my literature review led me to expect that confidence interval misconceptions were largely the result of misunderstanding variation. Although it may be the case that high school students and college freshmen struggle with the idea of variation, I found that this was not the case for any of the pre-service teachers in my study. It is possible that students who are learning statistics for the first time have not developed expert conceptions about variation, and that influences many of their misconceptions as presumed in previous studies (e.g. Gauvrit & Morsanyi, 2014; Zhang & Stephens, 2016; Cobb et al., 2003; Garfield & Ben-Zvi, 2005; Konold & Pollatsek, 2002; Blanco, 2016). It is also possible that statistics educators are more likely to have developed expert conceptions of variation, yet lack an expert conception that would allow them to conceptualize an ASD as observed in this study. This may not be surprising given that variation itself is an intuitive concept manifest in everyday situations while the ASD is not manifest in real world experiences. For example, variation occurs over time in how long it takes to commute between home and work, in the price of a gallon of gas at a certain station, and in the number of goals scored by a soccer team. This could mean that students have existing p-prims about variation, and formal education serves to link those p-prims to causal nets relating to statistical inference.

CHAPTER 5: DISCUSSION

In this section I discuss the results of my study as they relate to my research questions. I will first summarize the misconceptions about confidence intervals that I observed existed within the group of pre-service teachers in this study. I will then summarize some of the experiences that may have led to those misconceptions. Last, I will summarize some inferences that led to changes in their conceptions.

First, prominent misconceptions included that: 1) a confidence interval allows us to make inferences about more than the mean of the population, 2) the range spanned by a confidence interval with a 95% confidence level is the range that contains 95% of the population data, 3) calculations for confidence intervals are not based on actual sampling distributions, and 4) statisticians must take multiple samples in order to compute a reliable confidence interval. The prominent misconceptions were considered to be problematic readouts or causal nets that were observed with more than one of the pre-service teachers in this study. Evidence of these misconceptions can be found in subsections 1a, 2a, 3a, and 4a of the results chapter.

Second, some experiences that may have led to these misconceptions include 1) a tendency to imagine normal distributions, 2) a prototypical image of sampling distributions 3) a class activity where students construct an MSD, and 4) theorems such as the Law of Large Numbers or Central Limit Theorem. The experiences were considered to involve centers of focus, or in other words, the visual, verbal, or conceptual objects that a student pays attention to (Lobato et. al., 2012). Greater elaboration about these experiences can be found in subsections 1b, 2b, 3b, and 4b of the results chapter.

Third, inferences that led to changes in student conceptions include 1) considering skewed population data, 2) considering the center and margins of error of confidence intervals,

3) considering the variability of sample means on a sampling distribution, and 4) comparing the mean of many samples to the mean of that same data in single sample. Inferences were defined as causal net systems of how an individual’s knowledge elements activate each other (Levrini & diSessa, 2008). Excerpts involving these inferences can be found in subsections 1c, 2c, 3c, and 4c of the results chapter.

Not Encapsulating the ASD Was A Cause of Confidence Interval Misconceptions

Essentially, the main problem was that pre-service teachers in this study did not encapsulate the ASD as an object rather than that they did not understand variation. This problem is clear, in part, from the prevalence of MSDs in their verbal justifications of confidence levels and intervals. I found it significant that at the beginning of the second interview 100% of the students in my study believed that statisticians collect multiple samples in order to calculate confidence intervals. Even after an introductory statistics course at the university and The Teaching and Learning of Statistics and Probability course, 0% of the students interviewed initially felt comfortable with the notion of constructing a confidence interval by taking only one sample. Instead, they felt that it was necessary to use an MSD. Activities such as the one by Pfaff and Weinberg (2009) may have been unsuccessful in helping students to conceptualize the confidence interval because they highlight “a large number of trials” which may encourage students to think about the MSD and its concomitant misconceptions. This is not wholly surprising since the repetition of many trials via simulation has been found to confuse some students about properties of the sampling distribution (Watkins, Bargagliotti, & Franklin, 2014). Hence, when students see simulations of sampling distributions while studying confidence intervals, they could be noticing the SDP and not encapsulating the ASD.

Suggestions for Future Research

It may be the case that most students do not realize that with each iteration of a new mean to the MSD or with each individual increase to n, the associated mean could jump to either side of the true mean. I wonder whether as students are talking about getting “closer and closer” to μ they see it as a one sided limit. Many students may not realize that increasing the data points that are factored into computing the mean may in fact result in an estimate that is further from the unknown parameter. While increasing sample size does decrease the variation of $\bar{\bar{x}}$ about μ, $\bar{\bar{x}}$ does not continuously approach μ as n approaches infinity. One cannot get far enough along in the process of increasing n and be assured that they are within a certain error range of μ because subsequent iterations of calculating a mean with additional data points could actually increase the error. It may be that students have an inappropriate link between the behavior of limits and the behavior of $\bar{\bar{x}}$ if they have studied the formal definition of a limit in calculus or the end behaviors of functions in algebra. Specifically, because of experience with f(x) approaching the limit L as x approaches infinity, students may think that as n increases, $\bar{\bar{x}}$ approaches μ. Investigating the prevalence of this misconception could be an interesting topic for future research.

Additionally, it may be the case that the phrase *all possible samples* is confused for some students. This phrase is often used when defining the ASD. However, *all possible samples* could mean all the samples that are possible to take with unlimited time and money or it could mean all the samples that are possible given real life restrictions to our sampling capacity. It may be an interesting topic for future study to learn whether students consider all possible samples to be all the infinitely many possible samples or all the samples that one might possibly take in real life.

Limitations and Directions for Future Research

The results of this study are not generalizable because participants were not randomly selected. Due to the nature of interviewing human subjects, it was necessary for all the pre-service teachers to consent to be interviewed, so this limitation was unavoidable. To minimize the impact of lurking variables associated with the type of student who may volunteer to participate in the study I selected a subset of students who consented to participate in the study. My goal in selecting which pre-service teachers to invite to participate in the study was to get a wide range of types of misconceptions manifest on the pretest. I did not attempt to invite equal numbers of males or females to participate because for the purposes of this study it was appropriate for the demographics of the participants to be similar to the demographics of the students enrolled in the Teaching and Learning of Statistics course.

Another limitation of this study was that I had to infer what the conceptions of the students in my study were and also infer the impact that potential centers of focus had on their conceptions. Researchers often face this limitation since it is difficult to identify what individuals are actually thinking. The inferences that I made stemmed specifically from what the pre-service teachers said and did during the interviews, and the pre-service teacher accounts as well as my inferences may or may not be accurate.

A third limitation of my study is that I could not definitively determine the extent to which the MSD influenced the pre-service teachers' conceptions. Part of this limitation was due to the fact that I did not want to ask leading questions since I did not want the participants to be cued as to whether the thoughts they were sharing with me were correct or incorrect ways to reason about confidence interval estimation. I concluded that the MSD is an existing conception largely based on the students' pattern of recommending that multiple samples need to be taken before statisticians can produce a credible confidence interval, on their use of illustrations of bell curves with many sample means plotted inside of them, and on their language which dealt with the sampling distribution as a tool for making approximations of the population mean. However, based on the information that the participants volunteered during the interviews I could not tell whether student conceptions included the idea that statisticians build an MSD then set the upper and lower boundaries of the confidence interval so as to encompass the middle CL% of all the sampled means. This may be an interesting question for future research.

A fourth limitation of my study is that I am not able to say that the activities the students were engaged in during the interviews resolved their misconceptions. Part of the issue with this limitation is that we cannot say that students would not continue using the misconception in the future since expert conceptions are not stable and individuals may resort to misconceptions or expert conceptions in different contexts. Future research may investigate concept projections of each of the misconceptions manifest in this study to see whether the misconceptions are more prevalent in certain contexts.

A fifth limitation of this study is that the participants all came from the same class which further limits the generalizations that can be appropriately made concerning the findings. However, it could be interesting to learn whether the activities taking place when students in this study were able to construct more expert conceptions are activities that motivate resolutions for other students with similar misconceptions. If a study is done in the future to investigate the utility of these activities for improving student conceptions, then that study should include a large, random sample to incorporate more diversity and to report on whether findings of this study may be appropriately generalized.

Conclusion

In this study I set out to learn about conceptions held by pre-service teachers concerning confidence intervals. I found that participants in this study had some misconceptions about confidence intervals. Four of those misconceptions were documented in this paper. First, some participants had a tendency to assume a normal population distribution which lead them to overgeneralize their inferences. Second, some participants believed that confidence intervals contain the middle CL% of the data. Third, all participants thought that an ASD does not have utility in calculating a confidence interval because it cannot be constructed in real life situations. Fourth, all students thought that in practice, statisticians must take many samples to build an MSD to approximate an ASD. Evidence of the misconceptions was documented with connections to classroom experiences that influenced the participant's misconceptions and with resolutions I observed during interviews as the participants moved from their misconceptions toward expert conceptions.

REFERENCES

Andrade, L., & Fernández, F. (2016). Interpretation of confidence interval facing the conflict. *Universal Journal of Educational Research*, 2687-2700.

APA. (2001). *Publication manual of the American psychological association* (5th ed.). Washington, DC: American Psychological Association.

Batanero, C., Burrill, G., & Reading, C. (2011). Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study: the 18th ICMI study. New York, NY: Springer.

Batanero, C., Tauber, L. M., & Sánchez, V. (2004). Students' reasoning about the normal distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 257–276). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Blanco, T. L. G. (2016). Statistics Education of Elementary Teachers: Pre-service Elementary Teachers' Statistical Reasoning and Misconceptions. University of Wyoming, Laramie, WY.

Castro-Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98-113.

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cobb, P., McClain, K., & Gravemeijer, K. P. E. (2003). Learning about statistical covariation. *Cognition and Instruction*, *21*, 1–78.

Conference Board of the Mathematical Sciences. (2001). The mathematical education of teachers. Providence, R.I. and Washington, D.C.: American Mathematical Society and Mathematical Association of America. http://cbmsweb.org/MET2/met2.pdf

Coulson, M., Fidler, F. & Cumming, G. (2005). Understanding of confidence intervals by researchers in psychology, behavioural neuroscience, and medicine. Manuscript in preparation.

Confrey, J. (1990). A review of the research on student conceptions in mathematics, science, and programming. *Review of Research in Education*, *16*(1), 3-56.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170-180.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*, 199–311.

Davis, R., Vinner, S (1986). The notion of limit: Some seemingly unavoidable misconception stages. Journal of Mathematical Behavior, *5*(3), 281-303.

De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2009). Intro Stats, Instructor's Edition, Third Edition, Pearson.

delMas, R. C., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55–82.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman, & P. B. Pufall (Eds.), *Constructivism in the computer age* (pp. 49-70). Mahwah, NJ: Lawrence Erlbaum Publishers.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, 1155-1191.

Feldon, D. F. (2010). Why magic bullets don't work. *Change: The Magazine of Higher Learning*, *42*(2), 15-21. doi: 10.1080/00091380903563043

Fidler, F. (2006). Should psychology abandon p-values and teach CIs instead? Evidence-based reforms in statistics education. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics* (pp. 1-6). Voorburg, The Netherlands: International Statistical Institute.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–123.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (GAISE) report: A preK-12 curriculum framework. Alexandria, VA: American Statistical Association.

Franklin, C. A., Bargagliotti, A. E., Case, C. A., Kader, G. D., Scheaffer, R. L., & Spangler, D. A. (2015). The statistical education of teachers. Alexandria, VA: American Statistical Association.

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92–99.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*(3), 372-396.

Gauvrit, N., & Morsanyi, K. (2014). The equiprobability bias from a mathematical and psychological perspective. *Advances in Cognitive Psychology*, *10*(4), 119.

Gilliland, D., & Vince Melfi, V. (2017). A note on confidence interval estimation and margin of error. *Journal of Statistics Education*, *18*(1), 1-8. doi:10.1080/10691898.2010.11889474

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology, 45*(3), 135-140.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38*(5), 427–437.

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

Hannula, M. S., Pehkonen, E., Maijala, H., & Soro, R. (2006). Levels of students' understanding on infinity. *Teaching Mathematics and Computer Science*, *4*(2), 317-337.

Harlow, L. L. (1997). Significance testing in introduction and overview. In L. L. Harlow, S. A. Muliak and J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 1-14). Mahwah, NJ: Lawrence Erlbaum Associates.

Healey, J. F. (2005). *Statistics. A tool for social research* (7th ed.). Belmont, CA: Thomson Wadsworth.

Hiebert, J. (2013). The constantly underestimated challenge of improving mathematics education. In K. R. Leatham (Ed.), *Vital direction for mathematics education research* (pp. 45-56). New York: Springer Science+Business Media.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematical knowledge for teaching. *Elementary School Journal*, *105*(1), 11-30.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*(1), 69-88.

Jones, S. R. (2015). Calculus limits involving infinity: The role of students' informal dynamic reasoning. *International Journal of Mathematics Education in Science and Technology*, *46*(1), 105-126.

Jones, S. R., Lim, Y., & Chandler, K. R. (2017). Teaching integration: How certain instructional moves may undermine the potential conceptual value of the Riemann sum and the Riemann integral. *International Journal of Science and Mathematics Education*, *15*(6), 1075-1095.

Kalinowski, P. (2010, July). Identifying misconceptions about confidence intervals. Proceedings of the eighth international conference on teaching statistics. IASE, Lijbljana, Slovenia, Refereed paper.

Kolar, V. M., & Čadež, T. H. (2012). Analysis of factors influencing the understanding of the concept of infinity. *Educational Studies in Mathematics*, *80*(3), 389-412.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Lane, D. M. (2014). Online Statistics Education: An Interactive Multimedia Course of Study. Retrieved from http://onlinestatbook.com

Levrini, O., & diSessa, A. A. (2008). How students learn from multiple contexts and definitions: Proper time as a coordination class. *Physical Review Special Topics: Physics Education Research*, *4*(1). doi: http://dx.doi.org/10.1103/PhysRevSTPER.4.010107

Lobato, J., Rhodehamel, B., & Hohensee, C. (2012). “Noticing” as an alternative transfer of learning process. *Journal of the Learning Sciences*, *21*(3), 433-482. DOI:10.1080/10508406.2012.682189

Lui, Y. (2005). Teachers' understandings of probability and statistical inference and their implications for professional development. Vanderbilt University, Nashville, TN.

Lyons, L. (2017). *Gallup Poll, May 2017*. Gallup World Inc. Retrieved from: http://www.gallup.com/poll/210902/amid-record-suffering-economy-improving-ukrainians.aspx?utm_source=genericbutton&utm_medium=organic&utm_campaign=sharing

McCarthy, J. (2017). *Gallup Poll, June 2017*. Gallup Economy Inc. Retrieved from: http://www.gallup.com/poll/211688/consumer-spending-stable-may-104.aspx?utm_source=genericbutton&utm_medium=organic&utm_campaign=sharing

Morsanyi, K., Primi, C., Chiesi, F., & Handley, S. (2009). The effects and side-effects of statistics education: Psychology students' (mis-)conceptions of probability. *Contemporary Educational Psychology*, *34*(3), 210-220.

Moore, D. S., & McCabe, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: W.H. Freeman and Company.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics* (6th ed.). New York: W.H. Freeman and Company.

Murtaugh, P. A. (2014). In defense of P values. *Ecology*, *95*(3), 611-617.

Norman, J. (2017). *Gallup Poll, May 2017*. Gallup Politics Inc. Retrieved from: http://www.gallup.com/poll/210917/views-moral-values-slip-seven-year-lows.aspx?utm_source=genericbutton&utm_medium=organic&utm_campaign=sharing

Pfaff, T. J., & Weinberg, A. (2009). Do hands-on activities increase student understanding?: A case study. *Journal of Statistics Education*, 17(3), 1-38.

Piaget, J. (1952). *The child's conception of number*. London: Routledge and Kegan Paul.

Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. London: Routledge and Kegan Paul.

Resnick, L. B. (1987). Constructing knowledge in school. In L. S. Liben (Ed.), *Development and learning: Conflict or congruence?* (pp. 19-50). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1009). Charlotte, NC: Information Age Publishing.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Simon, M. (2006). Key developmental understandings in mathematics: A direction for investigating and establishing learning goals. *Mathematical Thinking and Learning, 8*, 359–371.

Silverman, J., & Thompson, P. W. (2008). Toward a framework for the development of mathematical knowledge for teaching. *Journal of Mathematics Teacher Education*, *11*(6), 499–511.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, *3*(2), 115-163.

Thaden-Koch, T. C., Dufresne, R. J., Gerace, W. J., Mestre, J. P., & Leonard, W. J. (2004). A coordination class analysis of judgments about animated motion. In *AIP Conference Proceedings* (Vol. 720, No. 1, pp. 57-60). Madison, WI.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician,* 70(2), 129-133.

Watkins, A. E., Bargagliotti, A., & Franklin, C. (2014). Simulation of the sampling distribution of the mean can mislead. *Journal of Statistics Education*, *22*(3), 1-20.

Vallecillos, A. (2000). Understanding of the logic of hypothesis testing amongst university students. *Journal for Didactics of Mathematics*, *21*, 101–123.

Vinner, S. (1997). The pseudo-conceptual and the pseudo-analytical thought processes in mathematics learning. *Educational Studies in Mathematics*, *34*(2), 97-129.

von Glaserfeld, E. (1987). Learning as a constructive activity. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zhang, Q., & Stephens, M. (2016). Teacher capacity as a key element of national curriculum reform in statistical thinking: A comparative study between Australia and China. In D. Ben-zvi and K. Makar (Eds.), *The teaching and learning of statistics* (pp. 301-313). Cham, Switzerland: Springer.

APPENDIX A: PRETEST QUESTIONS

Please respond to each question even if you are unsure of whether your answer is correct.

1) What does a confidence interval mean?

2) Can you provide an example of a confidence interval in the real world? (You may make up the numbers and situation.)

3) What factors influence the width of a confidence interval? Please describe how the factors you have identified influence the width of a confidence interval.

Please circle a number that corresponds to your level of conviction for each of the previous responses. Higher numbers are associated with higher levels of conviction.

	1: I bet this response is wrong 5: I know my answer is right				
Question 1	1	2	3	4	5
Question 2	1	2	3	4	5
Question 3	1	2	3	4	5

The following questions all apply to the scenario described below.

A researcher is studying the sleeping habits of college freshmen. She collects data from a random sample of 100 freshmen to estimate the mean number of hours that freshmen get at night, and calculates a confidence interval of (6.8, 8.1) using a 90% confidence interval.

Would the confidence interval be larger, smaller, or the same if the researcher had used a confidence level of 95% instead of 90%? Please justify your answer.

True or false: The purpose of the confidence interval above is to determine the mean for the sample of 100 freshmen. Please justify your answer.

Would the confidence interval be larger, smaller, or the same if the researcher had used a sample of 500 freshmen instead of 100? Please justify your answer.

True or false: Because the level of confidence is 90%, then the confidence interval (6.8, 8.1) gives a range of the number of hours of sleep that 90% of all the freshmen get per night. Please justify your answer.

APPENDIX B: INTERVIEW PROTOCOL

Interview One Questions

The interviewer will ask the student “how did you decide that,” “how do you know,” and “how sure are you” throughout the interview as appropriate. Interview questions will be reworded by the interviewer when clarifications are deemed necessary.

How are confidence intervals constructed?

Here is a copy of your pretest. Will you please explain your thinking to me for each of the responses?

How does the width of a 90% confidence interval compare to the width of a 95% confidence interval for the same data?

What does a confidence interval give us plausible values for?

Are sample sizes and the widths of confidence intervals related?

What percent of the population data lies inside of a confidence interval when the level of confidence is 95%?

What is your prior experience with learning confidence intervals?

Interview Two Questions

The initial questions for interview two resumed having students discuss responses to their pretests. Depending on whether each student response aligns with a misconception or with an expert conception, one of the following two sets of follow up questions will be used.

Set 1 – When a misconception is evident

Can you draw a picture to demonstrate what you are thinking?

What metaphors help you to understand this concept?

Can you think of another example for which the answer is similar?

What if we changed the confidence level? The sample size? Took a new sample?

Set 2 – When an expert conception is evident

Why does that work?

How are you thinking about this problem differently than you did during the pretest?

What experiences were you having when you began thinking about this differently? For example, were you asking certain questions, having a group discussion, or doing empirical examples? Can you elaborate?

What questions do you have related to this problem?

Following a review of the student pretests, I asked students to design their own hypothetical study. I consistently investigated conceptions regarding the number of samples that are necessary to take. Each student was prompted to describe the process of going from sample data to calculating a confidence interval and to justify any related theorems.