STUDENT CONCEPTUALIZATION OF THE INTERPRETATION OF THE CONFIDENCE INTERVAL AND THE CONFIDENCE LEVEL: IDENTIFYING SIMILARITIES AND DIFFERENCES IN STUDENT CONCEPT IMAGES OF CONFIDENCE INTERVALS

by

KRISTEN E. ROLAND

(Under the Direction of Jennifer J. Kaplan and Amy Ellis)

ABSTRACT

A world beyond p < 0.05 requires researchers to be mindful of reporting their statistical results. Confidence intervals are one of several ways to improve communication of inferential results. While initial research of the (mis)conceptions of confidence intervals published researchers and students hold exists, little research has focused on the cognitive development required to understand confidence intervals robustly. This dissertation study aimed to identify similarities and differences among aspects of individuals' concept image of confidence intervals, focusing on the interpretation of confidence intervals and interpretation of confidence levels.

Initial concept images (Tall & Vinner, 1981) and developmental clouds (Thompson et al., 2014) for the concept of confidence intervals were developed to guide the creation of the interview protocols used in this dissertation study. Participants took part in task-based interviews focused on conceptualizations of the interpretation of confidence intervals and the interpretation of confidence levels. A thematic analysis of eleven participants' interviews was conducted using the hypothesized concept images and methodology proposed by Powell et al. (2003).

In addition to confirmation of (mis)conceptions reported in the literature, three new conceptualizations of the interpretations were found: 1) using the capture/not capture explanation or 2) using the long-run interpretation of the confidence level as the interpretation of the confidence interval, and 3) discussing the connection between the confidence level and the width of the confidence interval. While five dimensions of an individuals' concept image of confidence intervals were identified as aspects of similarities and differences that may lead to different conceptualizations of interpretations, this dissertation study focused on two: 1) the word confident and 2) the concept of the confidence level. Themes among the participants' conceptualizations of confident and confidence level were identified. Frameworks were proposed for classifying conceptualizations of the interpretations, the word confident, the concept of confidence level, and the concept of confidence interval.

Future research is needed to identify productive and nonproductive paths for development of a robust concept image for confidence intervals. Implications for teaching include a proposed method for introducing confidence intervals that focuses on developing the concept of coverage probability.

INDEX WORDS: Statistics Education, Confidence Intervals, Interpretation of Confidence Interval, Interpretation of Confidence Level, Confident, Confidence Level, Concept Image, Developmental Cloud

STUDENT CONCEPTUALIZATION OF THE INTERPRETATION OF THE CONFIDENCE INTERVAL AND THE CONFIDENCE LEVEL: IDENTIFYING SIMILARITIES AND DIFFERENCES IN STUDENT CONCEPT IMAGES OF CONFIDENCE INTERVALS

by

KRISTEN E. ROLAND

B. S., Sonoma State University, 2009

M. S., University of Rhode Island, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

STUDENT CONCEPTUALIZATION OF THE INTERPRETATION OF THE CONFIDENCE INTERVAL AND THE CONFIDENCE LEVEL: IDENTIFYING SIMILARITIES AND DIFFERENCES IN STUDENT CONCEPT IMAGES OF CONFIDENCE INTERVALS

by

KRISTEN E. ROLAND

Major Professor:	Jennifer J. Kaplan
Co-Major Professor:	Amy Ellis
Committee:	AnnaMarie Conner
	Julie Luft

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
August 2020

DEDICATION

In Loving Memory of

Dr. and Mrs. William C. Howrie Jr.

Mr. and Mrs. Charles F. Roland Jr.

"Be good. Do well in school."

Poppi, I did it!

They say *it takes a village* to grow and to accomplish great things. This dissertation and my graduate school success would not have been possible without the strength, love and support of the many people in my life. I dedicate this work to all of you.

To my *Village*:

Charles and Diane Roland

Evan, Noel, Ginny, Annabeth, and Karla

Sheri Johnson

Krista Varanyak

Britton, Kathy, Laurel, Lisa, and Mallory

ACKNOWLEDGEMENTS

First, I wish to thank Jennifer J. Kaplan for the support and guidance you have given me over the course of the many years we have worked together. I would not be the statistics education researcher I am today without your mentorship. Thank you for all of the editing and multiple reads of this document – that alone shows dedication and strength of support! Of course, future readers will greatly appreciate your editing support, as well! Thank you to Amy Ellis. Your mentorship through this process has helped me in unimaginable ways and my current and future education research will forever be better because of the guidance and support you provided. To AnnaMarie Conner and Julie Luft: Thank you for the feedback and thoughtful suggestions throughout this dissertation process. This document is infinitely better because of your guidance.

A big thank you to my mathematics education family! Katrina and Pamela, thank you for all of your help with scheduling and rescheduling rooms for the many hours of interviews. This dissertation study would not have been possible without that data. I will forever be grateful to the many faculty who mentored and supported my development in education research. Jonathan, Julia, Brandon, Halil, Biyao, and everyone else – thank you for all of the support over the years.

Finally, I would like to acknowledge and thank the many participants in this dissertation study. The field of statistics education will forever be changed because of your willingness to take part in this study.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

News reporting of data and models reached unprecedented levels in 2020, the year that this dissertation was written. The world is responding to the coronavirus pandemic, producing models of disease spread, containment methods, and the effectiveness of treatments and vaccines for public consumption. Also, as the United States approaches the 2020 presidential election, polling results will become prolific in the media. Consuming this information requires global citizens to be knowledgeable consumers of data, scientific results, and statistical inference results – in other words be statistically literate citizens. *Statistical literacy* refers to a person's ability to 1) interpret and critically evaluate, and 2) discuss or communicate statistical information and data-related arguments (Gal, 2002). Interpreting and communicating the uncertainty in modeling and in inferential results, such as confidence intervals, can often be lost in translation. For example, when the results of the 2016 election did not match point estimates produced from forecasts produced by sites such as FiveThirtyEight, many citizens claimed the polls and modeling process had been wrong (Krall, 2016). The reality is the election results were within the error bands of the forecasts (see in particular the electoral college counts forecast at https://projects.fivethirtyeight.com/2016-election-forecast/#odds), something a citizen with a strong understanding of confidence intervals would know. Communicating the uncertainty associated with confidence intervals requires both the producer and consumer of the information to have robust understanding of the interpretation of the confidence interval and the interpretation of the confidence level. This dissertation study aims to expose the knowledge required to communicate inferential results using confidence intervals.

Over the past two decades, statistical inference, particularly its use in published research, has been criticized by researchers and research communities(Wasserstein et al., 2019a). The root of the issue was not statistical inference itself, but rather *how* researchers use statistical inference and *how* researchers interpret their results. The American Psychological Association (APA) funded a taskforce to understand better how statistical inference, particularly hypothesis testing, was being used in published research. The recommendations from this taskforce changed the guidelines for the publication of inference results in research. The APA recommended the use of confidence intervals as an efficient way to promote precision, significance, and understanding (APA, 2001). The American Statistical Association (ASA) had a two-step approach to influencing change in reporting of statistical inference: 1) publication of a formal statement about the use of p-values in published research (Wasserstein & Lazar, 2016) and 2) a special issue of *The American Statistician* with recommendations for implementation of the formal statement (see: Wasserstein et al., 2019b; Wasserstein & Lazar, 2016). The first document argued only for an increased understanding of statistical inference (Wasserstein & Lazar, 2016), but the special issue made stronger recommendations (Wasserstein et al., 2019b). The authors of the special issue encouraged readers to imagine a world beyond $p < 0.05$, a world without benchmarks and “statistically significant” results. Instead, in this world, the authors argued, researchers would ATOM: “**A**ccept uncertainty. Be **t**houghtful, **o**pen, and **m**odest” (Wasserstein et al., 2019a, p. 2) in their reporting of results. This suggested change in reporting will only come to fruition if researchers learn to be thoughtful in their communication of inferential results. Statisticians and statistics education researchers realize this change requires collaboration with many research communities, development of new curricula designed to help students develop understanding of statistical reasoning and uncertainty (Steel et al., 2019), and development of examples of good communication of the same aspects of statistical inference (Goodman, 2019). Reporting of confidence intervals and effect sizes, advocated by both the APA and the ASA, is one way to accept uncertainty and be modest in reporting results. The recommendations, for example the reporting of interval estimates and elimination of the word *significant* from scientific reporting (Fricker et al., 2019; Matthews, 2019; Wasserstein et al., 2019a), should lead to an increased use of confidence intervals in scientific research.

Confidence intervals, which can be considered *intervals of plausible values for an unknown value of a parameter*, provide an alternative to the conclusions drawn from hypothesis testing. Hypothesis tests require the researcher to provide additional information, such as effect size, for the reader to gain an understanding of the practical significance of the findings. By reporting confidence intervals, readers not only have an interval of plausible values for a parameter, they also can infer the sampling variability associated with the statistic through the width of the interval, a more transparent way of understanding the practical significance of the findings than a *p*-value. When interpreting confidence intervals, Wasserstein et al. suggested that researchers:

> include (1) discussing both the upper and lower limits and whether they have different practical implications, (2) paying no particular attention to whether the interval includes the null value, and (3) remembering that an interval is itself an

> estimate subject to error and generally provides only a rough indications of uncertainty given that all of the assumptions used to create it are correct, and, thus, for example, does not ‘rule out’ values outside the interval (2019a, p. 6).

Studies that explored the knowledge retained about interpretations of confidence intervals (e.g., Belia et al., 2005; Crooks et al., 2019) and interpretations of the conclusion and *p*-values from hypothesis testing (see: Castro Sotos et al., 2007), illuminated what they call misconceptions (and this document will call (mis)conceptions, see *(Mis)conceptions Reconceived: Epistemological Perspective*, but little work had been done to understand how these (mis)conceptions form. Several researchers have attempted to study how learners interpret confidence intervals, with mixed results (e.g., Andrade et al., 2014; Andrade & Fernández, 2016; Fidler, 2005; Grant & Nathan, 2008; Henriques, 2016; Kalinowski et al., 2018). These studies focused on a diverse group of students (pre-service teachers, psychology undergraduate and graduate students, and general population undergraduate students) at different levels of statistical experience, with the aim of identifying common (mis)conceptions their subjects had about confidence intervals.

Complicating matters, the idea of confidence intervals is a highly complex set of concepts within statistical inference. In the process of condensing these complex concepts and connections for instruction, epistemological and pedagogical decisions have been made by textbook authors and instructors (called didactical transpositions, see: Chevallard & Bosch, 2014) to create a delicate balance between 1) the mathematically and statistically intensive aspects that form the basis for many statistical concepts and 2) mathematical and statistical maturity of enrolled students. This balancing act could mean a sacrifice of depth of knowledge and mathematical support. This, in turn, could result in creating epistemological obstacles: knowledge obstacles fostered as a consequence of the decisions instructors make that truncate complex and contextual knowledge in order to meet their students' current cognitive level (Brousseau, 1997). While impossible to avoid all obstacles, a review of current publications indicated little work in the statistics education research field has studied these obstacles. Furthermore, little work has focused on understanding the connections that students have among the statistical concepts required for a robust understanding of confidence intervals. To help learners better understand statistical inference, work needs to be done to understand the concepts and connections that individuals have that about statistical inference.

Problem Statement

Published research about how individuals understand confidence intervals was sparked, in part, by the internal debate among researchers of the APA concerning the use of hypothesis testing (NHST) over confidence intervals (CIs) in published works (Wilkinson, 1999). Following recommendations in the 5th edition of the APA publication manual (APA, 2001) to include confidence intervals and remove hypothesis testing from published reports, researchers began to study how published researchers, particularly in the fields of psychology, medical neuroscience, and medicine, came to interpret confidence intervals (e.g., Belia et al., 2005; G. Cumming et al., 2004; G. Cumming & Finch, 2005). These studies indicated that most researchers did not understand confidence intervals fully. Within the statistics education field, research focused on student (mis)conceptions of the interpretation of confidence intervals (e.g., Crooks, 2014; Fidler, 2006) and on aids for instruction (e.g., Bertie & Farrington, 2003; Gordon & Gordon, 2020; Hagtvedt et al., 2008). From these works, researchers identified several (mis)conceptions individuals held about confidence intervals:

1. the confidence level represents the probability that the sample mean is in the interval (e.g., Crooks, 2014; J. Cumming et al., 2014),
2. the confidence level represents the percentage of potential values of the statistic within the interval (i.e. that for a 95% confidence interval for a population mean, 95% of sample means fell within the given interval) (e.g., Crooks, 2014; Kalinowski, 2010),
3. the confidence interval represents the range in which the confidence level percentage of individual scores were (e.g., Crooks, 2014; J. Cumming et al., 2014),
4. the confidence interval is a fixed interval with a random value of the parameter (e.g., Andrade & Fernández, 2016; Grant & Nathan, 2008).

While the field identified possible (mis)conceptions individuals hold, there is a lack of understanding of the *concept image*, defined as "the cognitive structure in the individual's mind that is associated with a given concept" (Tall & Vinner, 1981, p. 151), that formed these (mis)conceptions of the concept of confidence intervals. Fidler (2005) and Crooks (2014) developed ideas of the knowledge (termed understanding in the papers) students need to understand confidence intervals. Broadly, these could be separated into two types of knowledge: *definitional* and *relational* knowledge (Fidler, 2005). *Definitional* knowledge was defined as the knowledge that a confidence interval is an estimate for an unknown value of a population parameter and is part of the inferential family (Fidler, 2005). This could include knowledge of the definition of the terms confidence interval and confidence level and knowledge of how to interpret the confidence interval (Crooks, 2014). *Relational* knowledge was defined as knowledge of the relationship among the components of a confidence interval (i.e. the relationship between confidence level and interval width) (Crooks, 2014; Fidler, 2005). Crooks (2014) also included the knowledge of "the distinction between samples and populations and how they are related" (Crooks, 2014, p. 15). While these two types of knowledge are part of the knowledge required for a robust understanding of confidence intervals, behind these definitional and relational characteristics lie a web of concepts that have yet to be explored. This study proposes to begin to identify the concept images that are developed in individuals with productive conceptions and in individuals with non-productive conceptions of confidence intervals (documented (mis)conceptions of confidence intervals).

Statement of Purpose and Research Questions

The primary focus of the current body of research about confidence intervals is on (mis)conceptions. The purpose of this study is to begin to explore the concept image that individuals develop about the concept of confidence intervals. Specifically, this study explores possible dimensions of the concept image of the concept of confidence intervals required for a robust understanding of the interpretation confidence intervals and confidence levels. Specifically, I am exploring the following questions:

1. What conceptualizations of interpretations of confidence intervals and interpretations of confidence levels do undergraduate and graduate students have?

2. What similarities and differences exist in undergraduate and graduate students' concept images of the concept of confidence intervals when they conceptualize interpretations of confidence intervals and interpretations of confidence levels?

These research questions begin to fill the gap in the literature about how students develop different conceptions of confidence intervals. I conducted a series of in-depth task-based clinical interviews designed to uncover the concept image students formed about confidence intervals, with participants from four levels of statistical coursework. These clinical interviews were designed to elicit student knowledge of the construction, definitional, and relational characteristics of confidence intervals. This study has both short-term and long-term implications. In the short term, this study helps future researchers understand what aspects of concept images developed by students are demonstrating similarities and differences in conception. By beginning to map the concept image of the concept of confidence intervals, future research can help identify the productive and non-productive pathways to robust understanding of confidence intervals. Thus, the long-term implications of this study, and future studies, should allow future instruction to help break the cycle of improper use and interpretation of confidence intervals. It is important to note that this research study is not intended to promote or refute the recommendations to abandon the p-value in favor of confidence intervals. Rather, this study intends to fill in the literature surrounding the known (mis)conceptions of the interpretations of confidence intervals, and interpretation of confidence levels.

CHAPTER 2

THEORETICAL FRAMEWORK AND LITERATURE REVIEW

The majority of literature on which this study was based consists of (mis)conception studies. The first section of this chapter discusses how this literature is viewed and how I distinguish between the epistemological assumptions of the literature and my epistemological beliefs. The following sections will focus on the theory of confidence intervals, the construct of didactical transposition and resulting obstacles, the potential epistemological obstacles within introductory statistics textbooks, the theoretical framework used in this study, the potential connections among concepts required for a robust understanding of confidence intervals, and the published (mis)conceptions individuals have about confidence intervals.

(Mis)conceptions Reconceived: Epistemological Perspective

(Mis)conception studies began in the early 1980's and became prevalent in the 1990's. Initially, the study of (mis)conceptions aided education researchers in bringing the attention of instructors to studies that identified possible reasons for students' misunderstanding of content (diSessa, 2006). This identification of (mis)conceptions strengthened the need for qualitative research designed to elicit understanding against the backdrop of quantitative studies (diSessa, 2006). On the other hand, (mis)conception studies were widely without theory development or testing and strongly emphasized the negative contributions of prior knowledge (diSessa, 2006).

Much of the prior research on the understanding and teaching of confidence intervals focused on (mis)conceptions demonstrated by published researchers and psychology students. To keep the intended tone of the authors of the studies reviewed in my dissertation, I used the word (mis)conceptions as the authors intended. My own opinion of (mis)conceptions, however, is that individuals form the '(mis)conceptions' through experiences in coursework, life, and research. In line with Smith et al. (1994), I believe focusing research on (mis)conceptions allows researchers to form deficit mindsets of their subjects, privilege the understanding of the researcher, and imply that (mis)conceptions can be eradicated. In this dissertation, both terms, conceptions and (mis)conceptions, are used. In general, the word (*mis)conception* is used to "designate a student conception that produces a systematic pattern of errors" (Smith et al., 1994, p. 119). In contrast, the word *conception* maintains a less negative connotation and could "identify and relate factors that students use to explain intriguing or problematic phenomena" (Smith et al., 1994, p. 119). Throughout this study, I used the term *conceptions* when discussing student knowledge and (mis)conceptions when discussing ideas from the published (mis)conception literature. I aim to avoid viewing my participants in a deficit manner, to identify the concept image participants were using when conceptualizing confidence intervals, and to not privilege my conception of the concept of confidence intervals over my participants'.

Definitions

While there was some continuity concerning constructs within statistics education, not all constructs were well defined. For ease of reading, here are a list of constructs and my definitions:

- *Standard deviation of the sampling distribution*: the theoretical standard deviation of the sampling distribution.
- *Standard error*: the estimated standard deviation of the sampling distribution. This can be either an approximated value from a simulated sampling distribution or a standard deviation of the sampling distribution calculated from a statistic.
- *Actualized/ Realized*: a descriptor to make the distinction between data that have been gathered and a theoretical sample that has yet to be selected. An actualized/ realized sample is one that has been collected.
- *Random Process*: the process through which random variables exist and are assigned probabilities. These can be considered models that are used to describe real-world data.
- *Interpretation of a Confidence Interval*: We are confidence-level% confident that the value of the parameter is between the upper and lower bounds of the interval.
- *Interpretation of a Confidence Level*: Approximately confidence-level% of all possible samples of size n from a population would produce confidence-level% confidence intervals that capture the true value of the parameter.
- *Confidence Level*: the term used to discuss the coverage probability of the confidence interval estimator.
- *Confident*: a word to refer to the difference between the calculated interval, which no longer has probability of capturing the unknown value of a parameter, and the random interval, which has probability associated with the random process through which the interval was constructed.

- *Estimator*: a function of a random variable(s) (i.e. point estimator, interval estimator).
- *Estimate*: the calculated value of an estimator calculated using an actualized collection of data.

A full list of definitions used in this document, including constructs, can be found in Appendix A.

Theory of Confidence Intervals

This sub-section discusses the general theory of confidence intervals and interpretations of confidence intervals, confidence levels and relational characteristics of confidence intervals. Confidence intervals are one aspect of statistical inference, the body of techniques through which a researcher attempts to learn about an unknown population based on information gathered from a sample.

General Theory of Confidence Intervals

Mathematical statistics textbooks typically begin the presentation of statistical inference with an introduction to the difference between parameters and statistics (point estimates), which includes how to select the best point estimator for an unknown value of a parameter (e.g., Casella & Berger, 2002; Chihara & Hesterberg, 2011; Wackerly et al., 2002). These textbooks then introduce the need to make inferences about the value of the unknown parameter while accounting for the "omnipresence of variability" (Cobb & Moore, 1997, p. 801), the foundation of statistics. This acknowledgement of variability leads to the discussion of better estimations for values of an unknown parameter, including interval estimators. One such type of interval estimators are *confidence intervals*, "interval estimators, together with a measure of confidence (usually a confidence coefficient)" (Casella & Berger, 2002, p. 419). The theory of confidence intervals discussed in this dissertation is frequentist, referring to the belief that probabilities represent long-run frequencies of repeated random experiments. Before moving forward with a discussion of the theoretical aspects of a confidence interval, several constructs are defined.

These definitions are based on those provided by Casella and Berger (2002):

- *Point Estimator*: a function, $T(x_1, x_2 \dots x_n)$, of a random variable that can be used to estimate an unknown value of a parameter, θ.
- *Point Estimate*: a realized value of a function of a random variable that can be used to estimate an unknown value of a parameter, θ.
    - *Note*: Ideally, this estimate was calculated from an estimator that is the *uniform minimum variance unbiased estimator* of the unknown value of a parameter. *Uniform minimum variance unbiased estimator* (UMVUE) ensured the estimator was unbiased (the expected value of the point estimator of θ equals θ for all possible values of θ) and the variance of the point estimator was less than or equal to the variance of all other possible unbiased point estimators.
- *Interval Estimator*: An interval estimator is an interval created from two functions of a random variable such that a random interval is created. Formally, given any pair of functions $L(x_1, \dots, x_n)$ and $U(x_1, \dots, x_n)$ of a sample that satisfy $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$, the random interval $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ is called an *interval estimator.*

- *Interval Estimate*: An interval estimate is a realized interval that was created from an interval estimator with a realized random sample that can be used to estimate a value of the unknown parameter, θ.
- *Coverage probability:* The probability that the random *interval estimator* covers the true value of the parameter θ. Formally: $P_\theta(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$ or $P([L(\boldsymbol{x}), U(\boldsymbol{x})]|\theta)$.
- *Confidence coefficient*: The infimum, the greatest lower bound of a set, of the coverage probabilities. Formally: $inf_\theta P_\theta(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$.

It is important to understand that the *interval estimator* is constructed from random elements from which probability is leveraged to formulate an interval that will contain the unknown value of the parameter, θ, referred to as the *coverage probability*. These statements do not describe a random parameter, θ, with a fixed interval. Thus, the coverage probability is the probability of the random interval estimator containing the value of the unknown parameter θ rather than the probability of the value of the unknown parameter is contained within the random interval (*interval estimate*). This distinction is difficult to imagine and understand, but is pivotal in understanding the interpretation of a confidence interval (Gilliland & Melfi, 2010).

A confidence interval can be defined from these constructs. Based on the definition of an interval estimator, we create the following general formula for a 100(1-α)% confidence interval:$P_\theta(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})]) = 1 - \alpha$. There are two generally accepted methods for calculating the values of $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$: inverting a likelihood ratio test (LRT) or using *pivots* (or *pivotal quantities*). Since LRTs can often be defined for all possible θ within a particular scenario, these can be a successful starting point to calculating a confidence interval. To invert a LRT, one inverts the acceptance region found to relate the probability that the random interval (or set) captures the fixed θ, formally defined as "for each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0: \theta = \theta_0$. For each $\boldsymbol{x} \in X$, define a set $C(\boldsymbol{x})$ in the parameter space by $C(\boldsymbol{x}) = \{\boldsymbol{\theta_0}: \boldsymbol{x} \in A(\theta_0)\}$" (Casella & Berger, 2002, p. 421). *Pivots* are functions of the random variable such that the distribution of the pivot is free from unknown parameters. Common pivots were: $Z = \frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$, $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, or $V = \frac{(n-1)\boldsymbol{S}^2}{\sigma^2} \sim \chi^2_{n-1}$. As an example of using pivots to create a confidence interval (example derived from Casella & Berger, 2002, p. 425), if we were investigating an independent and identically distributed $(X_1, \dots, X_n)$ sample that we believed came from a normally distributed population with unknown population mean, μ, and unknown population variance, σ^2, then we know that:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$P\left(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

$$P\left(-t_{\alpha/2,n-1} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{\alpha/2,n-1} * S/\sqrt{n} < \mu < \bar{X} + t_{\alpha/2,n-1} * S/\sqrt{n}\right) = 1 - \alpha$$

Then, we could construct the $100(1-\alpha)\%$ confidence interval to be:

$$\bar{X} \pm t_{n-1,\alpha/2} * \frac{S}{\sqrt{n}},$$

where $t_{n-1,\alpha/2}$ is the *confidence coefficient*.

It was important to note that there is a subtle difference between $\bar{X} \pm t_{n-1,{}^{\alpha}/_{2}} * \frac{s}{\sqrt{n}}$ prior to collecting data from a random variable and $\bar{X} \pm t_{n-1,{}^{\alpha}/_{2}} * \frac{s}{\sqrt{n}}$ after collecting data from a random variable. If $\bar{X}$ and *s* were functions of a random variable, then the endpoints of the interval would also be random. Thus,

$$P\left(\bar{X} - t_{{}^{\alpha}/_{2},n-1} * {}^{S}/_{\sqrt{n}} < \mu < \bar{X} + t_{{}^{\alpha}/_{2},n-1} * {}^{S}/_{\sqrt{n}}\right) = 1 - \alpha$$

was a true statement. If $\bar{X}$ and *s* are now fixed values from an actualized sample from a random variable, the statement:

$$P\left(\bar{X} - t_{{}^{\alpha}/_{2},n-1} * {}^{S}/_{\sqrt{n}} < \mu < \bar{X} + t_{{}^{\alpha}/_{2},n-1} * {}^{S}/_{\sqrt{n}}\right) = 1 - \alpha$$

is no longer true. Here, there are no longer any random values to which probability can be assigned. Instead, the statement is either true or false: the value of the parameter, μ, is either in the interval (P=1) or is not in the interval (P=0). Since the value of the parameter is unknown, the individual calculating the interval cannot know whether the interval is one of the $100(1-\alpha)\%$ that captures the value of the parameter or one of the $100(\alpha)\%$ that does not capture the value of the parameter. Therefore, there was confidence level *probability* that the value of the parameter would be within the random interval, but there is confidence level *confidence* that the value of the parameter is within the actualized interval. *Confidence* refers to the probability associated with the random process – not the probability that the value of unknown parameter is within the actualized interval.

As a general rule, the *confidence interval* contains two components: 1) the confidence coefficient and 2) either the standard error or the standard deviation of the sampling distribution. The combination of these two elements is the *margin of error* of the random interval. Although there is no formal definition of *margin of error* within

Casella and Berger (2002), it can be considered the maximum possible distance between the point estimate and the plausible value of the parameter while maintaining the determined level of confidence. The next section discusses the generally accepted forms of interpretations of the confidence interval and confidence level.

General Interpretations of Confidence Intervals

There are three components of a confidence interval that warrant interpretation: the confidence level, the confidence interval, and the relational characteristics of the confidence interval. The frequentist *interpretation of the confidence level* is generally agreed upon by the statistics field as: 1) approximately 100(1-α)% of all possible 100(1-α)% confidence intervals should capture the unknown value of the parameter and/ or 2) prior to collecting data, the probability of a 100(1-α)% confidence interval capturing the value of the unknown parameter was 100(1-α)%. For instance, we could say the random interval $\bar{X} \pm t_{n-1,\alpha/2} * \frac{S}{\sqrt{n}}$ would capture the fixed value of an unknown parameter 100(1-α)% of the time. Equivalently, we could say that 100(1-α)% of intervals created through this process would capture the unknown value of the parameter. These interpretations are fairly straightforward because both describe the coverage probability either in terms of long-run probability or the probability associated with a random sample from a random variable. The interpretation of the confidence interval is not as straightforward.

The theory described above indicates that the random endpoints are calculated based on the probability they surrounded the fixed but unknown value of the parameter. Once the interval estimator becomes an interval estimate there is no longer a random variable for which one could calculate a probability – everything within the probability statement is fixed. Therefore, the interpretation of the confidence interval is given in terms of the probability that the random interval would capture the value of the unknown parameter, rather than the probability that a specific interval captured the value of the unknown parameter. This leads to the commonly accepted *interpretation of the confidence interval*: “We are 100(1-α)% confident that the calculated interval contains the true value of the parameter.” *Confident* in this sentence implies that we are confident in the random process of sampling and calculating confidence intervals rather than the probability that any one actualized confidence interval contained the value of the unknown parameter. This interpretation is not without its own set of controversies. Morey, et al. identified three fallacies they believed exist within interpretations of confidence intervals:

1. *The Fundamental Confidence Fallacy* described as:“If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observed interval contains the true value is also X%; or, alternatively, we can have X% confidence that the observed interval contains the true value.”
2. *The Precision Fallacy*: “the width of a confidence interval indicates the precision of our knowledge about the [unknown value of the] parameter. Narrow confidence intervals correspond to precise knowledge, while wide confidence errors correspond to imprecise knowledge.”
3. *The Likelihood Fallacy*: “a confidence interval contains the likely values for the parameter. Values inside the confidence interval are more likely than those outside. This fallacy exists in several varieties, sometimes involving

> plausibility, credibility, or reasonableness of beliefs about the [unknown value of the] parameter." (2016, pp. 104–105)

Morey et al. argued that the frequentist assumptions of traditional (Neyman) confidence intervals do not lend themselves to interpretation once an actualized interval has been created. Instead, the authors argued for the use of Bayesian credible intervals. For the purpose of this dissertation, only frequentist confidence intervals were considered. While Morey et al. identified a difficult aspect of the interpretation of a confidence interval, the general consensus of the textbooks sampled, including the textbook for the introductory statistics course taught at the university where this dissertation study took place, appeared to side with some form of the interpretation: "We are 100(1-α)% confident that the calculated interval contains the true value of the parameter" (e.g., Agresti et al., 2017; De Veaux et al., 2018; Larson & Farber, 2019; McClave & Sincich, 2017; Triola & Iossi, 2018). Hoekstra, et al. (2018) discovered when reviewing textbooks for the fallacies, 93% (21 out of 23 textbooks) presented one of the fallacies: 61% of the textbooks reported the *Fundamental Confidence Fallacy* and the *Likelihood Fallacy* was found in 43% of the textbooks. Furthermore, Wasserstein et al. (2019a) appeared to acknowledge the difficulty in interpreting confidence intervals, but focused on the interpretation of a confidence interval by dichotomizing the results similar to hypothesis testing. The authors appeared to support the idea of interpreting the confidence interval as a "compatibility" interval. Amrhein et al. (2019) suggested the term compatibility because of the nature of the calculation of the interval – the interval is a result of a random process and the sample data, meaning the resulting interval represents the values of the parameter that are compatible with the sample data and coverage probability. Despite this controversy, the students participating in this research study learned *the interpretation of a confidence interval* to be of similar structure to: “We were 100(1-α)% confident that the calculated interval would contain the true value of the parameter.” Therefore, I will use this interpretation of a confidence interval in the analysis.

There are two types of *relational characteristics* for confidence intervals that statisticians should understand: 1) how sample size affects the width of the interval and 2) how the confidence level affects the width of the interval. There existed an inverse relationship between sample size and confidence width. As the sample size increases, the sampling variability (variability between-sample) decreases. This, in turn, decreases the standard deviation of the sampling distribution (or the estimated standard error), which causes the margin of error to decrease. As the confidence level increases, the width of the confidence interval increases, forming a direct relationship. The increase in width is a result of the increase in the range of the sampling distribution corresponding to the coverage probability, or the confidence level. These relationships inform study design decisions. In the next section, I will discuss how these interpretations were addressed in introductory statistics textbooks.

Confidence Intervals in Introductory Statistics

Introductory statistics textbooks have the difficult task of consolidating highly complex and connected statistical topics into an introductory course, often aimed at students without the mathematical maturity to understand the theoretical underpinnings of most statistical concepts. This phenomenon can be discussed using the constructs identified by Brousseau (1997): didactical transpositions and resulting obstacles. This section elaborates on these constructs and introduces examples from introductory statistics textbooks that could potentially cause obstacles for students to overcome as they learn more advanced statistics.

Didactical Transposition and Epistemological Obstacles

The issues with converting deeply connected and historically driven subject knowledge into knowledge taught in classrooms can be many and complex. The idea of transforming contextualized subject knowing into curricular knowledge can be called *didactical transposition* and can cause teachers and instructors to make decisions that could cause difficulties for students later (Brousseau, 1997). These difficulties can be considered *obstacles*, which Bachelard (as cited in Brousseau, 1997) originally defined as errors and failures connected to prior knowledge originally helpful, but that has become less helpful when applied to larger knowledge domains (Brousseau, 1997). An example is when students learn that multiplication makes numbers bigger when they learn about multiplication of natural numbers. These students may become confused when learning about multiplication of rational numbers where the product may not be bigger than the multiplicands. Similar to the idea of (mis)conceptions, these types of errors are persistent and not unexpected (Brousseau, 1997). Obstacles of this type are not just errors or (mis)conceptions. These obstacles, such as the learning multiplication and subtraction with natural numbers before learning multiplication with fractions and subtraction with negative numbers, are epistemologically necessary: allowing students access to curricular concepts within a limited domain that is appropriate for the child's current development that will become an obstacle to overcome at a later point when the domain is expanded.

Brousseau (1997) suggested methods for attending to these obstacles similar to those suggested by Smith et al. (1994) for addressing (mis)conceptions: the continual application of new situations that challenged the obstacle and caused the learner to reject it as a plausible piece of knowledge. There are three possible origins for obstacles: 1) ontogenic origin, 2) didactical origin, and 3) epistemological origin. *Obstacles of ontogenic origin* are produced from issues that arise from knowledge developed during different stages of a learner's development. A learner can develop ideas based on the neurophysiological limitations and cognitive abilities available to him/her at the time of development (Brousseau, 1997). In contrast, *obstacles of didactical origin* are based on the decisions that teachers and education systems made concerning curriculum or had socio-cultural origins (Brousseau, 1997). The final type of obstacles, *obstacles of epistemological origin*, are based on necessary decisions made during the didactical transposition of knowledge, which is often required for the development of knowledge (Brousseau, 1997). Epistemological Obstacles are necessary choices, such as the decision to introduce multiplication as repeated groups, that are inherent in instruction. It would be problematic to attempt to teach multiplication with rational or integer numbers prior to developing a child's understanding of rational or integer numbers. This epistemological decision exists in the second-grade curriculum when students can develop an understanding of multiplication through repeated grouping but only know natural numbers. Therefore, the obstacle of multiplication makes bigger is an understandable and necessary conception if a student does not have a conception of rational or integer numbers. It is important to realize that these types of obstacles cannot be ignored (Brousseau, 1997).

Obstacles of epistemological origin are faced by mathematics and statistics educators when they must transpose knowledge that is complex, highly contextualized, and historic in nature for introductory students. If instructors maintain the level of complexity and context required by mathematical and statistical content, students might not assimilate the information in the intended manner because the students do not have the ability to integrate the knowledge in the same manner as a statistician or mathematician. Brousseau described this event as “choosing to define the knowledge in its definitive form and ‘organization’ as a language with the risk that this language is not adapted to the students’ development” (1997, p. 111). On the other side, if educators define content with information that is less complex and less contextualized within the historical architecture of the subject, the instructor risks students formulating knowledge that is “incomplete knowledge … and produces obstacles which can be more or less overcome by the student and the teacher but which provokes many difficulties” (Brousseau, 1997, p. 111).

Didactical Transposition in Introductory Statistics

Gilliland and Melfi (2010) pointed out that recent studies on subjects’ understanding of confidence intervals did not provide completely accurate statements concerning confidence intervals and margins of error. By doing so, they identified epistemological and didactical obstacles, although they did not use these terms. Complicating introductory statistics instruction, these topics are introduced to students who do not have the required advanced mathematical background (often algebra is the only pre-requisite) using simplistic or trivial examples. The topics traditionally used to introduce confidence intervals are inferences for one population mean when the population standard deviation, σ, is known (e.g., Larson & Farber, 2019; McClave & Sincich, 2017; Weiss & Weiss, 2016), one population mean when σ is unknown, and/or one population proportion (e.g., Agresti et al., 2017; De Veaux et al., 2018; Triola & Iossi, 2018). It could be argued that the situation when σ is known is an artificial situation since inference about the population mean is redundant if one knows the true standard deviation of the population. In the other two situations, there are added complexities because we must also estimate the standard deviation of the sampling distribution (*standard error*). Gilliland and Melfi (2010) pointed out that ignoring the additional error in estimating both unknown values of the parameters can be detrimental to students who continue beyond an introductory level. The authors also pointed out that the statements described by many textbooks were not completely correct or accurate in describing these more complex situations and that this issue was amplified by researchers who focus solely on overly simplified topics as defined in introductory statistics without considering the complex nature of the topic.

The inference models used in introductory statistics courses are fairly straightforward in terms of calculation with pivot functions that transform the distribution of the variable into either a normally distributed or t-distributed variable, e.g. $Z = \frac{\bar{X}-\mu}{\sigma}$, $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$, or $Z = \frac{\hat{p}-p}{\sqrt{(\hat{p}\hat{q})/n}}$. The intervals created by these pivot functions are often referred to as one-population z and t intervals and allow the confidence intervals to be of a standard form: $\hat{\theta} \pm MoE$, where $\hat{\theta}$ is a point estimate for the value of the unknown parameter, θ, and *MoE* is the margin of error. With symmetric distributions, the margin of error is easier to identify. Centering instruction on the statement that all confidence intervals are of the form $\hat{\theta} \pm MoE$, however, is similar to the multiplication always makes bigger conception: an epistemological obstacle that would need to be overcome as students learn about non-symmetric distributions and confidence intervals.

Introductory statistics textbooks, courses, and researchers often use the interpretation that 100(1-α)% of sample statistics are expected to fall within +/- margin of error of the value of the unknown parameter. This was, however, a place of critique by Gilliland and Melfi (2010). When applied to situations when σ is known, this interpretation is correct. When we were using the standard error by approximating p or σ with $\hat{p}$ or s, however, the interpretation is incorrect. Two dissertations explored student understanding of the margin of error (Liu, 2005; Noll, 2007). Both of these studies were critiqued by Gilliland and Melfi (2010) for their limited example of margin of error by using trivial examples, when the population standard deviation was known, to explore student understanding. By knowing the population standard deviation, the exact standard deviation of the sampling distribution for sample means could be calculated. This means that the margin of error would be the same length for all confidence intervals. Therefore, the interpretation given in Thompson and Liu (2005) of "the margin of error ± 5% to mean '95% of sample statistics fall within ± 5% of the unknown population" was correct, but only for this situation. Once the population standard deviation is not known, which is usually the case, this interpretation is no longer correct. If the actual standard deviation of the sampling distribution is not known, then the margin of error would change based on the given sample. By teaching the overly simplified situation, Thompson, Liu and Noll might have created an unnecessary didactical obstacle for students to overcome in future instruction, or, worse, that these students would possess generally incorrect knowledge thinking it was true.

Gilliland and Melfi further demonstrated that the definitional and relational examples provided in introductory statistics courses are not suitable for situations when the coverage probability derived from the sampling distribution is problematic (Gilliland & Melfi, 2010). They also critiqued the simplification of interpretations of confidence intervals and margins of error presented in textbooks such as: 1) De Veaux, Velleman, and Bock's *Intro Stat* (as referenced in: Gilliland & Melfi, 2010), which incorrectly interprets a margin of error as the proportion of *all* random samples that will be within the *estimated* margin of error of the true proportion (emphasis added) and 2) Noll's and Liu's dissertation studies, which limited discussion of margin of error to the trivial example, using the true standard deviation of the true sampling distribution. Gilliland and Melfi identified these issues as problematic in part because students could develop incorrect understanding of the process in the methods used to compute and interpret confidence intervals.

Review of Literature

Most of the research concerning the (mis)conceptions of people of all statistical experiences about confidence intervals had been published in fields of study other than statistics education. This section of this chapter will focus on (mis)conceptions published within existing literature, broken into three sections based on (mis)conception type, followed by a section on primary methods for teaching confidence intervals.

Confidence Intervals

Initially, the research on conceptions of confidence intervals attended to those held by published researchers. These studies focused on conceptions of confidence intervals held by supposed experts as evidence of how misunderstood confidence intervals were by practicing quantitative researchers (see: Belia et al., 2005; Fidler, 2005; Hoekstra et al., 2014). Education researchers had begun to compare novice knowledge to expert knowledge by including comparisons of introductory statistics students' responses to those provided by quantitative researchers and/or faculty members. While the populations for the studies varied, most of the published literature researched (mis)conceptions of psychology undergraduate and graduate students at different levels of statistical experience. The populations and levels of statistical experience have been summarized in Table 1. Most studies used a fairly large sample with closed form assessments (Canal & Ruiz, 2015; Coulson et al., 2010; Crooks et al., 2019; J. Cumming et al., 2014; Fidler, 2005; García-Pérez & Alcalá-Quintana, 2016; Henriques, 2016; Hoekstra et al., 2012, 2014; Kalinowski, 2010; Kalinowski et al., 2018), while several were smaller sample case study reports (Andrade & Fernández, 2016; Canal & Ruiz, 2015; Grant & Nathan, 2008).

As there was little qualitative follow-up to most of these studies, there were few recommendations in the published literature for instruction (for exceptions see: Andrade et al., 2014; Andrade & Fernández, 2016; Canal & Ruiz, 2015) or how these (mis)conceptions developed (for exceptions see: Grant & Nathan, 2008; Liu, 2005; Noll, 2007; Thompson & Liu, 2005). It is important to note that there was no evidence in the above studies that the documented (mis)conceptions were restricted to particular levels of statistical experience. In the few expert/novice comparison studies, similar (mis)conceptions were present at both levels (Canal & Gutiérrez, 2010; Crooks et al., 2019; Hoekstra et al., 2014).

Table 1

List of Published Works of Confidence Interval (Mis)conceptions

Population	Level	Citation	Sample Size	Study Type
Psychology	Undergraduate Students	Crooks (2014)	80	Mixed Methods
		Crooks et al. (2019)	40	Mixed Methods
		Fidler (2005)	180	Closed Form
		García-Pérez and Alcalá-Quintana (2016)	313	Closed Form
		Hoekstra et al. (2014)	596	Closed Form
	Graduate Students	Crooks et al. (2019)	40	Mixed Methods
		Fidler (2005)	180	Closed Form
		García-Pérez and Alcalá-Quintana (2016)	313	Closed Form
		Grant and Nathan (2008)	3	Qualitative Interviews
		Hoekstra et al. (2012)	62	Open Form Questions
Behavior Neuroscience, Psychology, and Medicine	Published researchers	Belia et al (2005)	473	Web-based Applet
		Coulson et al (2010)	473	Web-based Applet
Psychology, Science and Medicine	Undergraduate and Graduate Students	Kalinowski (2010)	473	Web-based Applet
		Kalinowski et al. (2018)	101/ 24	Mixed Methods
Pre-service or In-service Teachers Mathematics Teachers		Andrade et al. (2014)	N/A	Teaching Suggestions
		Andrade and Fernández (2016)	3	Teaching Experiment
		Canal and Ruiz (2015)	297	Closed Form
		Foster (2014)	12	Open Form
		Liu (2005)	8	Teaching Experiments
		Noll (2007)	68/5	Mixed Methods
		Thompson and Liu (2005)	8	Teaching Experiment
General	Undergraduate Students	Canal and Gutiérrez (2010)	15	Open Form
		J. Cumming et al. (2014)	710/ 207	Open Form
		Fidler (2005)	180	Closed Form
		Henriques (2016)	33	Mixed Form

The literature presented in this section will be separated based on their results. There are three broad categories for the documented (mis)conceptions: definitional, relational, and comparison of hypothesis tests and confidence intervals. Definitional studies have results focused on the definitional knowledge (interpretation of confidence intervals, interpretation of confidence levels, procedural knowledge of the calculation of confidence intervals, and the randomness associated with confidence intervals) required to understand confidence intervals. Relational studies reported findings about the effects changes in sample size, confidence level had on the width of the confidence interval. Comparison of hypothesis tests and confidence intervals studies reported results about differences in interpretation and conclusions drawn when individuals are presented with hypothesis test and/or confidence interval results.

Definitional Characteristics.

Definitional characteristics of confidence intervals can be categorized as: 1) procedurally computing the confidence interval (see: Henriques, 2016), 2) interpreting the confidence level, and 3) interpreting the confidence interval. In the foundational study, Fidler (2005) reported that 180 undergraduate psychology students from various statistics experiences (1 to 4 semesters of statistics) selected one of the following items from a list of definitions of a confidence interval (choices as appear in the paper, percentages are proportion of students who selected the statement as the correct interpretation, note only a is correct): a) the plausible values for the population mean (22%), b) the plausible values for the sample mean (38%), c) the range of individual scores (8%), and d) the range of individual scores within one standard deviation (20%), with 21% responding unsure. The (mis)conception concerning the sample mean was further investigated using another closed form question: “The 95% confidence interval has a ____% chance of capturing the sample mean” (Fidler, 2005, p. 212). Eighty-four percent of students responded with 95%, instead of the correct 100%. Fidler stated that confusion between sample and population may affect student understanding of confidence intervals. These findings became the basis for most future studies. For ease of reading, I have separated the further findings based on the categories of the definitional characteristics.

Interpreting the Confidence Level

When *interpreting the confidence level* of a confidence interval, Canal and Gutiérrez (2010) found that only 50% of experts (professionals devoted to statistics or its teaching and senior statistics students) and 36% of students (active engineering or business administration) correctly identified the statement "If 200 [confidence intervals] of the same process were generated, approximately 10 of them will not contain the population mean," which is the correct interpretation of the 95% confidence level. Documented misinterpretations of the confidence level were 1) the proportion of the population data contained within the interval (Canal & Gutiérrez, 2010; delMas et al., 2007), 2) the probability the unknown value of the parameter was in the interval (Andrade et al., 2014; Andrade & Fernández, 2016; Foster, 2014; Hoekstra et al., 2014), 3) the proportion of sample statistics (or replication statistics) in the interval (delMas et al., 2007; Kalinowski et al., 2018), 4) the probability the statistic was in the interval (Canal & Gutiérrez, 2010), and 5) related to accuracy (Canal & Ruiz, 2015; Kalinowski et al., 2018). There were three general themes to these (mis)conceptions: issues with parameters, issues with statistics, and issues with population data.

The *proportion of the population contained in the interval* (mis)conception was explored in two studies. Canal and Gutiérrez (2010) presented a novice/expert comparison study in which they gave their participants a series of true/false statements with interpretations of the confidence level[1]. On this questionnaire, two statements were presented to participants: 1) "95% of weights are between 42 and 48 pounds" and 2) "Most weights are between 42 and 48 pounds" (Canal & Gutiérrez, 2010, p. 2), both of which are incorrect. The majority of experts disagreed with the first statement, but approximately half of them agreed with the second. Approximately 50% and 85% of students agreed with statement 1 and statement 2, respectively. The belief that the confidence level is the proportion of the population within the confidence interval was affirmed by delMas et al. (2007). delMas and colleagues found that less than one-third of the participants identified the statement implying the confidence level was the proportion of the population within the confidence interval as invalid. The authors did find that at post-test, two-thirds of participants identified this statement as invalid.

Andrade and colleagues (2014; 2016) designed instructional tasks to be used with pre-service teachers (PSTs) to attempt to disrupt the (mis)conception that the confidence level is the probability that the unknown value of the parameter is within the limits of the confidence interval using the students' understanding of probability and sample space. In both studies, the authors found that the PSTs did not view the confidence interval as an event that was part of a larger sample space and the instructional tasks were unable to change the PSTs assumptions that the confidence level was the probability the value of the unknown parameter was in the interval. Foster (2014) administered a short questionnaire to colleagues to identify possible lexically ambiguous differences in

[1] Note: The authors of the article referred to their statements as definitions of the confidence interval. This literature review made a distinction between statements concerning the interpretation of the confidence interval and the confidence level. This change is evident of my choice in categorizing statements according to their intended meaning.

sentences used to interpret confidence intervals. He provided his participants with the following sentences:

> A. About 95% of the time the true population mean lies inside the confidence interval.
>
> B. I'm about 95% sure that the confidence interval contains the true population mean.
>
> C. The probability that the true population mean is within the confidence interval is 95%.
>
> D. There is a 95% chance that the true population mean is inside the confidence interval. (Foster, 2014, p. 27)

Out of the 12 participants, all but one viewed statement A as acceptable or were ambivalent about its correctness. Most (9) considered statement B as unacceptable. The remaining two statements were mostly considered as no opinion/ambivalent about the correctness of the statement. Canal and Gutiérrez found that 40.4% of their experts and 56.2% of their novices identified the following statement as true: "The probability that the interval includes the population mean is 95%" (2010, p. 2). Hoekstra et al. found that approximately 50% of participants believed the following statement to be a correct interpretation of the given confidence interval: "There is a 95% probability that the true mean lies between 0.1 and 0.4" (2014, p. 1160). Even more problematic, 66% of first year students, 79% of masters students and 58% of researchers believed: "If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4" (Hoekstra et al., 2014, p. 1160). These findings were slightly controversial (see: García-Pérez & Alcalá-Quintana, 2016; Miller & Ulrich, 2016) due to the study design and the authors' critique of interpretations of confidence intervals. Regardless, it is disconcerting that so many participants endorsed these two statements.

There were two general types of (mis)conceptions with respect to the confidence level and the sample/statistic: the confidence level is 1) the proportion of sample statistics or replication statistics (for the difference between sample statistic and replication statistic see: G. Cumming et al., 2004; G. Cumming & Maillardet, 2006) within the limits of the confidence interval and 2) the probability the sample statistic is within the interval. Similar to Fidler (2005), the statement "the probability that the interval includes the sample mean is 95%" was selected as a true statement by one-third of experts and just over half of students in Canal and Gutiérrez's (2010) study. delMas et al. (2007) found that the percentage of students who chose this as a valid statement increased from pre-test to post-test. Additionally, delMas et al. found that a majority of students identified the confidence level as the proportion of statistics that would be within the confidence interval limits.

The final set of (mis)conceptions concerning the confidence level was discussed by Canal and Ruiz (2015) and Kalinowski et al. (2018). Students incorrectly believed the confidence level and accuracy of the interval were related, where accuracy meant the likelihood of the unknown value of the parameter being within the interval. Kalinowski et al. documented students notions that lower levels of confidence (i.e. 50%) indicated a lack of data and implied that the unknown value of the parameter "could fall anywhere" (2018, p. 11). Canal and Ruiz (2015) found that students believed the probability the unknown value of the parameter was in the interval corresponded to the accuracy of the confidence interval. Before moving on to the (mis)conceptions concerning the interpretation of a confidence interval, it is important to note that these studies generally reported responses to closed-response questionnaires without further exploration into the ideas behind these statements.

Interpreting a Confidence Interval

Fidler (2005) addressed most of the (mis)conceptions around the *interpretation of a confidence interval*: range of plausible values for the statistic and range of individual scores. Out of 180 undergraduate psychology students from various statistics experience (1 to 4 semesters of statistics), 38% selected the plausible values for the sample mean response, and 8% selected the range of individual scores response (Fidler, 2005). Kalinowski et al. (2018) explored student understanding of the relative likelihood of each point within the interval as the unknown value of the parameter. The authors suggested that this likelihood was not equal across the interval, but rather was more likely near the center of the interval and less likely towards the end of the interval. They referred to the likelihood as *a cat's eye* (for further information see: G. Cumming & Finch, 2005; Kalinowski et al., 2018). While conducting qualitative interviews, the authors affirmed that students held the (mis)conceptions found in Fidler (2005). Only a few studies have discussed correct interpretations of definitional characteristics. Fidler (2005) found that 22% of participants correctly identified the confidence interval as a range of plausible values for the population mean. delMas et al. (2007) reported that half of the students recognized the valid interpretation of a confidence interval on pre-test, which increased to three-quarters on post-test. delMas et al. were concerned that while the percentage of students who correctly identified valid interpretations increased at post-test, many students indicated as valid both a correct and an incorrect interpretation of confidence intervals, even at post-test. While Hoekstra et al. (2014) disagreed that the statement "we can be 95% confident that the true mean lies between 0.1 and 0.4" was a correct interpretation of a confidence interval (see fallacy descriptions above in the *Theory of Confidence Intervals* section), they found that 49% of first year students, 50% of masters students, and 55% of researchers in their study endorsed this statement as a true statement.

Grant and Nathan (2008) extended Fidler's (2005) findings using the theory of conceptual metaphors to identify productive and non-productive conceptual metaphors. The authors used these metaphors to determine participants' understanding of *fixed interval*. The authors defined the productive metaphor to be *changing ring metaphor*, which implies that confidence intervals are changing rings around a fixed point. The non-productive metaphor was the *fixed disk metaphor*, which implied that the confidence intervals are change points on a fixed disk. This was a non-productive metaphor because it led students to develop the conception that the confidence interval was fixed rather than a random variable, when it was the value of parameter that was fixed. Using students in a graduate statistical methods for social sciences course, Grant and Nathan found that these two metaphors existed and often existed together in the minds of the participants. Of the three participants, two maintained the fixed disk metaphor, and one held a changing ring metaphor but used statements classified as a fixed disk metaphor to contrast her changing ring metaphor answers.

Foster (2014) made the point in his discussion of the interpretations of confidence intervals that the language of the interpretation is rather precise and could be lexically ambiguous. He used the two statements: "the population mean lies within the confidence interval" and "the confidence interval contains the population mean" to demonstrate one possible confusion. In the first statement, there is implied action on the population mean – indicating that the population mean is not a fixed, but unknown, quantity. The second statement removes the action of the population mean. It is generally accepted that the first statement indicates a student thought the interval was fixed but the value of the parameter was changing. The second statement was considered correct. Foster likened these statements to: "the milk is in the fridge" and "the fridge contains the milk," which are equivalent statements in English and argued that language needed to be more precise.

Callaert (2007) commented on the difficulty surrounding the use of the word "confident" in interpreting the confidence interval. From his experiences at AP readings, he found that students had learned that confident was the correct word to use but questioned whether students knew what this word meant. Based on discussions the author had with other AP readers, the other instructors expressed concern about the blanket teaching of "confident" indicating that students and teachers alike may read "95% confident" and "95% probability" as the same (Callaert, 2007). Kaplan et al. (2010) conducted a more rigorous study consisting of student responses defining the word confident as part of a larger study about lexically ambiguous statistical terms. The authors reported that the most frequent category contained student responses that "mentioned a level of certainty or surety" (Kaplan et al., 2010, p. 10), which was broken into several sub-themes: 1) about the location of a value, 2) of something vague or unspecified, and 3) that something is correct. Based on a pilot study of 49 students and a random sample of 100 students in the validation study, Kaplan et al. classified the students' definition of confident as presented in Table 2. The authors reported that a large proportion of students defined confident using a response classified as a level of surety or certainty and a smaller proportion defined confident implying a high level of certainty. Kaplan et al. were troubled by the large proportion of students with incoherent meanings. From these three studies, the interpretation of a confidence interval and the meaning of the word confident are not as straightforward as it would appear.

Table 2

Student Definitions of Confidence

Definition		Number of Subjects	
		Pilot Study	Validation Sample
A level of surety or certainty	about the location of a value	1 (2%)	6 (9%)
	of something vague or unspecified	12 (26%)	23 (35%)
	that something is correct	5 (11%)	4 (6%)
Have a high level of certainty, be very sure		9 (19%)	7 (11%)
Accuracy or precision		1 (2%)	5 (8%)
An interval		6 (13%)	1 (2%)
Ability to provide evidence		1 (2%)	1 (2%)
Not classified		12 (26%)	17 (26%)

Note. From “Lexical Ambiguity in Statistics: How students use and define the words: association, average, confidence, random and spread,” J. J. Kaplan, D. G. Fisher, N. T. Rogness, 2010. *Journal of Statistics Education*, *18*(2), p. 11 (https://doi.org/10.1080/10691898.2010.11889491). 2010 by Jennifer Kaplan, Dianne G. Fisher, and Neal T. Rogness. Reprinted with permission.

Relational Characteristics

There were two main relational characteristics identified in the literature: 1) how changes in confidence level and 2) how changes in sample size affect the width of a confident interval. In the foundational study, Fidler (2005) found that students thought: 1) confidence interval width increased with sample size (20% of participants), 2) confidence interval width was unaffected by sample size (29% of participants), and 3) a 90%

confidence interval was wider than a 95% confidence interval (for the same data) (73% of participants). Fidler also found that 16% of students were able to identify correctly that the confidence interval width decreased with sample size. Canal and Gutiérrez (2010) followed up with a series of true and false statements looking at 1) the relationship between a fixed sample size and a wider confidence interval if the confidence level increases (83% of experts and 51% of students correctly identified this as a false statement), 2) the relationship between a fixed confidence level and a narrower confidence interval if we increase the sample size (64% of experts and 52% of students correctly identified this as a false statement), and 3) the relationship between increasing the population standard deviation and decreasing confidence interval width (increase in population standard deviation would increase the standard deviation of the sampling distribution, which increased the confidence interval width; 9% of experts and 43% of students agreed with the statement). Kalinowski and colleagues (2010; 2018) observed that students held the belief that a 95% confidence interval was approximately double in length to a 50% confidence interval. Kalinowski et al. (2018) suggested that their use of *the cat's eye* figures reduced these relational misconceptions. I, however, hypothesize that increased connections to sampling distributions will help students better conceptualize these relationships.

Comparison of Understanding of Hypothesis Tests and Confidence Intervals

Belia, et al. (2005) were the first to investigate how published researchers in psychology, behavioral neuroscience, and medicine interpreted graphs of confidence intervals and standard error bar results visually. Graphical displays were recommended by the APA as the best method for displaying statistical results. Specifically, the subjects were asked to identify, using sliding confidence intervals, the location of statistically significant differences between two groups at a significance level of 0.05. The authors found that the respondents placed confidence intervals at an average distance that was equivalent to a *p*-value of 0.009 but were too lenient on standard error. The subjects' response to the standard error context had an average distance equivalent to a *p*-value of 0.109. They also found that respondents were unable to determine the difference between confidence interval and standard error bars even though the distinction would require different visual rules (for "rules of thumb" see: G. Cumming, 2007). Lastly, the respondents were unable to determine the correct relationship for a repeated measures example.

Moving from this foundational piece, Coulson et al. (2010) investigated how these same population of researchers were able to interpret study findings when given the results in the form of null hypothesis statistical testing or confidence intervals. The authors determined that neither hypothesis testing nor confidence intervals improved the researchers' understanding of the hypothetical study findings. For both of these studies, however, the method for collecting participants was to send emails to published researchers in top journals. As a result, response bias in these studies could underestimate the actual amount of (mis)conceptions held by researchers.

Hoekstra et al. (2012) also explored the connection between interpretations of results when presented as null hypothesis testing or provided as confidence intervals. In an activity sent via email to PhD students, participants were presented with eight scenarios: four scenarios with results presented with confidence intervals and four scenarios with results presented with hypothesis tests. The authors found that participants appeared to be more sure of the correct decision when scenarios were presented with results from a hypothesis test while participants were more capable of interpreting results when presented with results in the form of confidence intervals. They found that participants were, on average, more sure that population effects existed when presented with results from a hypothesis test than with confidence intervals. There were fewer inferential mistakes, fewer references to significance, and more references to effect size when presented with confidence intervals over hypothesis testing (Hoekstra et al., 2012, p. 1049). From these findings, the researchers concluded that confidence intervals improve interpretations of results but do not guarantee against misinterpretations.

Primary Methods for Teaching Confidence Intervals

While it is beneficial to know the difficulties associated with understanding of confidence intervals, little research had attempted to study how to improve instruction on confidence intervals (exceptions: Andrade et al., 2014; Andrade & Fernández, 2016; Crooks, 2014; Inzunza, 2018). Only a few studies identified ways to improve student conceptions focusing on a particular (mis)conception. Andrade and Fernández (2016) attempted to create a task to address the *fixed interval* (mis)conception, but were unsuccessful in changing the deep rooted prior-conceptions the students held. Crooks (2014) attempted to address conceptual understanding of confidence intervals through alternating procedural and conceptual instruction, but was unable to demonstrate particular improvement.

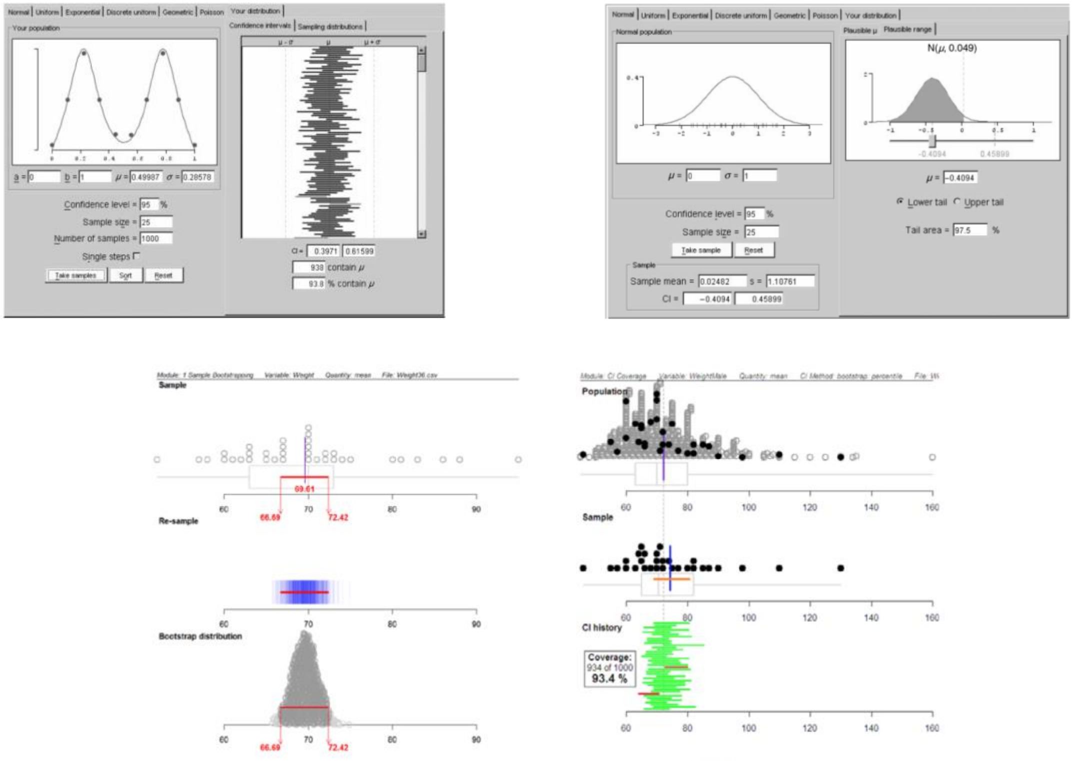
Over the course of the past two decades, the availability of technology in classrooms has changed the way statistics is taught. This is evident from Recommendation 5: Use technology to explore concepts and analyze data in the *GAISE:*

College Report 2016 (2017). In addition, statistical methods such as bootstrap techniques or simulation-based techniques broadened the curricula available to instructors of introductory statistics. Curricula, such as *Statistical Thinking: A Simulation Approach to Modeling Uncertainty* (Zieffler & Catalysts for Change, 2018), *Introduction to Statistical Investigations* (Tintle, Chance, Cobb, Rossman, et al., 2015), and *StatKey* (Lock et al., 2013) help instructors teach statistics using a simulation-based or bootstrap-based inference approaches, which allows students to visualize the abstract and theoretical aspects of statistical inference (J. Cumming et al., 2014).

Visualization has become an important part of statistics education. Confidence intervals are no exception. Many applets have been developed to create visualizations of confidence intervals (e.g., Bertie & Farrington, 2003; Hagtvedt et al., 2007, 2008; Mills, 2002; Tintle, Chance, Cobb, Roy, et al., 2015; Williams & Williams, 2018). Applets for confidence intervals typically display two types of scenarios. The first type of applet allows students to visualize the *confidence level* interpretation. They allow the user to take multiple samples from a population and display the confidence interval corresponding to each sample. The user should see that approximately *confidence level%* of the *confidence level%* confidence intervals capture the true value of the population parameter. These applets are similar to those in Figure 1a, which was one of the applets discussed in Bertie and Farrington (2003). Figure 1b displays an applet unique to Bertie and Farrington (2003). The authors designed this applet to help students visualize the *plausible values* interpretation of a confidence interval as the range of plausible values for the actual value of the population parameter. In this applet, the user draws a sample from the population and moves the distribution on the right side of the applet to identify plausible values for the population parameter (μ in this example).

Figure 1

Examples of Applets for Confidence Intervals



a = 0 b = 1 μ = 0.49987 σ = 0.28578

Confidence level = 95 %

Sample size = 25

Number of samples = 1000

Single steps □

CI = 0.3971 0.61599

938 contain μ

93.8 % contain μ

μ = 0 σ = 1

Confidence level = 95 %

Sample size = 25

Sample mean = 0.02482 s = 1.10761

CI = −0.4094 0.45899

μ = −0.4094

● Lower tail ○ Upper tail

Tail area = 97.5 %

Module: 1 Sample Bootstrapping Variable: Weight Quantity: mean File: Weight36.csv

Figure 1. *VIT* bootstrap confidence interval module

Module: CI Coverage Variable: WeightMale Quantity: mean CI Method: bootstrap: percentile File: W

Figure 2. *VIT* bootstrap confidence interval coverage module

Note. A and B: From "Teaching Confidence Intervals with Java Applets" by A. Bertie and P. Farrington, 2003, *Teaching Statistics*. *25*(3), p. 72-73 (https://doi.org/10.1111/1467-9639.00134). Copyright 2003 by John Wiley and Sons. Reprinted with permission. C: From "Using Bootstrap Dynamic Visualizations in Teaching" by J. Cumming, C. Miller, and M. Pfannkuch, in K. Makar and R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)* (p. 3), 2014, International Statistics Institute. Copyright 2014 by the Ninth International Conference on Teaching Statistics (ICOTS9). Reprinted with permission.

An alternative type of applet is used in bootstrap or simulation-based curricula for introductory statistics. One such applet is provided in Cumming et al, (2014) (see Figure 1c), which combines the displays for the sample, re-sample, and bootstrap plots and the *confidence level* interpretation applet. While these applets have been shown to improve some naïve conceptions, they have not been able to address all prior conceptions that students hold. Cumming, et al. (2014) concluded their article stating that a possible limitation of these applets and teaching methods is the constraints often applied to the teaching of these topics. The authors had only two lessons to introduce this statistical method to students. Instead, the authors suggested that students be exposed to these topics for a prolonged period of time.

Textbooks also use the relatively simple z- and t-intervals to address the more complex side of confidence intervals: interpreting the confidence interval, the confidence level, and the margin of error (e.g., Agresti et al., 2017; De Veaux et al., 2018; Triola & Iossi, 2018; Weiss & Weiss, 2016). The interpretation of a confidence interval is often presented as a sentence that explains, in context, the confidence interval constructed. This often includes discussing the confidence level in terms of the long-run probability the interval will contain the actual value of the parameter (e.g., Agresti et al., 2017; McClave & Sincich, 2017; Rossman & Chance, 2012; Sullivan, 2013; Triola & Iossi, 2018). Some textbooks explain what the term confident means in statistics: confident in the random process used to construct the intervals (e.g., McClave & Sincich, 2017; Moore & Notz, 2009; Rossman & Chance, 2012; Triola & Iossi, 2018; Utts, 2005). Some also explain the difference between confident and probability by explaining the effect the change in word choice has on the interpretation: confidence in the random processes versus probability the constructed interval captured or did not capture the value of the parameter (e.g., De Veaux et al., 2018; Lock et al., 2013; Peck, 2014; Triola & Iossi, 2018). If the margin of error was discussed as more than the value added and subtracted to the point estimate to create the confidence interval, it was defined as the maximum possible distance between the point estimate and the value of the parameter (e.g., Agresti et al., 2017; Larson & Farber, 2019; Peck, 2014; Triola & Iossi, 2018; Weiss & Weiss, 2016). De Veaux et al. (2018), Lock et al. (2013), Peck (2014), and Triola (2018) describe the difference between random and fixed values.

There were two relational characteristics of confidence intervals typically stated as universally true in an introductory statistics textbooks: 1) the change in the confidence level is directly related to the change in interval width (as the confidence level increases, the width of the interval increases and vice versa), 2) the change in the sample size while maintaining all other features of the sample constant is inversely related to the change in the width of the interval (as the sample size increases the width of the confidence level decreases) (e.g., Agresti et al., 2017; De Veaux et al., 2018; Triola & Iossi, 2018; Weiss & Weiss, 2016). These relational situations are easy to demonstrate visually and describe mathematically when using z- and t-intervals. Mathematically, it is fairly straightforward to demonstrate how changes in sample size change the standard deviation of the sampling distribution or the standard error and changing the confidence level changes the confidence coefficient, which changes the width of the interval. This can be done without connecting the formulas to sampling distributions, allowing for a missed connection for students. There are many applets that demonstrate visually the effects of sample size on the sampling distribution for $\bar{x}$ or $\hat{p}$ (see for example: Rossman Chance Applets, http://www.rossmanchance.com/applets/OneSample.html, and Online Stat Book, http://onlinestatbook.com/stat_sim/sampling_dist/).

Summary of Conceptions

In summary, there have been several studies that identified (mis)conceptions about the definitional and relational characteristics of confidence intervals and the comparison of null hypothesis testing and confidence intervals. Most studies were conducted using psychology students and faculty. Expert studies were mostly conducted with researchers in psychology, behavioral neuroscience, and medicine. Pre-service teachers were identified to hold similar conceptions as psychology students, but little work had been done with general population students. The conceptions found throughout the literature has been summarized in Table 3.

Table 3

Summary of the (Mis)conceptions Found in the Literature

Type	Category	Issue with	Name	Explanation	References
Definitional	Confidence Level	Population	Proportion of the Population	The confidence level is the proportion of the population within the confidence interval limits.	Canal and Gutiérrez (2010); del Mas et al. (2007)
		Parameter	Probability of the value of the parameter within the interval	The confidence level is the probability the value of the parameter is within the confidence interval limits.	Andrade et al. (2014); Andrade et al. (2016); Foster (2014); Hoekstra et al. (2014)
			Accuracy	The confidence level indicates a lack of data/ lack of accuracy.	Kalinowski et al. (2018)
		Statistic	Proportion of Sample Statistics	The confidence level is the proportion of the sample statistics within the confidence interval limits.	del Mas et al. (2007); Kalinowski et al. (2018)
			Probability of statistic within the interval	The confidence level is the probability the interval includes the sample mean.	Canal and Gutiérrez (2010); Fidler (2005)
	Confidence Interval	Population	Range of individual scores	The interval is the range of the population data.	Fidler (2005)
		Statistic	Range of values for the statistic	The interval is the range of plausible values for the statistic of interest.	Fidler (2005)
		Fixed Interval	Fixed Interval	The interval is either a fixed disk or a changing ring	Grant and Nathan (2008)
Relational		Confidence Level	Confidence Level	Changes in confidence level have a direct relationship with confidence interval width	Canal and Gutiérrez (2010); Canal and Ruiz (2015); Fidler (2005); Kalinowski (2010); Kalinowski et al (2018)
		Sample Size	Sample Size	Changes in sample size have an inverse relationship with confidence interval width	Canal and Gutiérrez (2010); Canal and Ruiz (2015); Fidler (2005); Kalinowski et al (2018)
		Standard Deviation	Standard Deviation	Changes in standard deviation have a direct relationship with confidence interval width	Canal and Gutiérrez (2010)
Comparison				Comparisons between confidence intervals and null hypothesis tests.	Belia et al. (2005); Coulson et al. (2010); Cumming et al. (2014); Fidler (2005); Hoekstra et al. (2012); Kalinowski et al. (2018)

Theoretical Framework

This section discusses the theoretical framework that is used in this study. There are two constructs that constitute the theoretical framework: concept image and developmental cloud. These constructs are defined in the first subsection. The next subsection discusses hypothesized concept images, using the structure of a developmental cloud, for the concept of confidence intervals, the concept of the interpretation of confidence intervals, and the concept of the interpretation of confidence level.

Concept Image and Developmental Cloud

A *concept image* is defined as:

> the total cognitive structure that is associated with the concept, which includes all the mental pictures and associated properties and processes. It is built up over the years through experiences of all kinds, changing as the individual meets new stimuli and matures. (Tall & Vinner, 1981, p. 152)

This definition resulted from the difficulty in "[distinguishing] between the mathematical concepts as formally defined and the cognitive processes by which they are conceived" (Tall & Vinner, 1981, p. 151). In particular, Tall and Vinner discussed the difference between formally defined mathematical concepts (*formal concept definition*) and an individual's definition of mathematical concepts (*personal concept definition*), particularly with respect to concepts that an individual may have encountered prior to formal definition or instruction. An individual's personal concept definition may deviate from a formal concept definition, and may vary because of a particular *evoked concept image*: "a portion of the concept image which is activated at a particular time" (Tall & Vinner, 1981, p. 152). These ideas are compatible with the obstacles identified earlier in this section. As an individual's personal concept definition evolves, experiences, such as early experiences with multiplication where multiplication makes bigger and subtraction where subtraction makes smaller, could lead to a different personal concept definition than the formal concept definition. This dissertation study aims to identify similarities and differences across individuals' concept image and personal concept definition for confidence intervals, interpretation for confidence intervals, and interpretation for confidence levels.

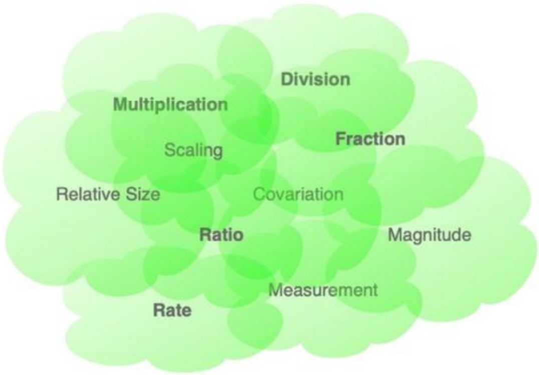
The second construct for this study is the *developmental cloud*, defined as "an ensemble of meanings and ways of thinking … entail[ing] some common ways of thinking while at the same time involving ways of thinking[2] that are unique to [the individual]" (Thompson et al., 2014, p. 14). Thompson et al. (2014) developed the idea of a developmental cloud from a need to describe complex ideas such as proportional reasoning and magnitude. To do this, the authors expanded Cobb and von Glasersfeld (1983, as cited in Thompson et al., 2014) and von Glaserfeld (1995, 1998 as cited in Thompson et al., 2014) interpretation of Piaget's concept of schemes. *Scheme*, as a construct to capture the complexity of thinking with higher-order mathematical concepts, was summarized by Thompson et al. as "an organization of action, operations, images, or [higher-order] schemes – which can have many entry points that trigger action and anticipations of outcomes of the organization's activity" (2014, p. 11). These schemes can "support flexible, innovative, creative thinking by making connections among

[2] The reader is referred to Thompson et al. (2014) for more information about the definition of "ways of thinking," and its difference from "understanding" and "meaning". These distinctions are not relevant to this study and are used interchangeably.

meanings and ways of thinking that are typical of high forms of thought" (Thompson et al., 2014, p. 12), which the authors considered groupings. By thinking of these schemes as groupings, it allowed Thompson et al. to consider the idea that students can "compose actions flexibly and keep in mind the different parts of their reasoning process" (Thompson et al., 2014, p. 12) and to develop the idea that schemes are developed both in a progression and in parallel. *Learning clouds* is the concept that Thompson et al. created to describe visually the idea that different processes are coordinated within a higher-order scheme, such as proportional reasoning or the concept of confidence interval. To reiterate a rather poignant statement in their article: "we hope to convey an image of parallel developments of ways of thinking that are always in interaction and yet constitute the span of those curricular concepts that compose proportional reasoning" (Thompson et al., 2014, p. 14). The developmental cloud is a way of visualizing of the coordination of many different meanings and ways of thinking to understand larger concepts. Figure 2 presents an example from Thompson et al. demonstrating the different meanings and ways of thinking that are coordinated within proportional reasoning. Each green cloud represents a learning cloud, or different aspects that are incorporated within proportional reasoning. These ideas of simultaneous development (or circular development) and developmental clouds were fundamental to the development of the concept image for confidence intervals, interpretation of confidence intervals, and interpretation of confidence levels that are presented in the next section.

Figure 2

Visual example of a developmental cloud



Note. From P. W. Thompson, M. P. Carlson, C. Byerley, and H. Hatfield, "Schemes for Thinking With Magnitudes: An Hypothesis About Foundational Reasoning Abilities in Algebra," in K. C. Moore, L. P. Steffe, and L. L. Hatfield (Eds.), *Epistemic Algebra Students: Emerging Models of Students' Algebraic Knowing* (WISDOMe Monographs, Vol. 4, p. 15), 2014, University of Wyoming. Copyright University of Wyoming 2014. Reprinted with permission.

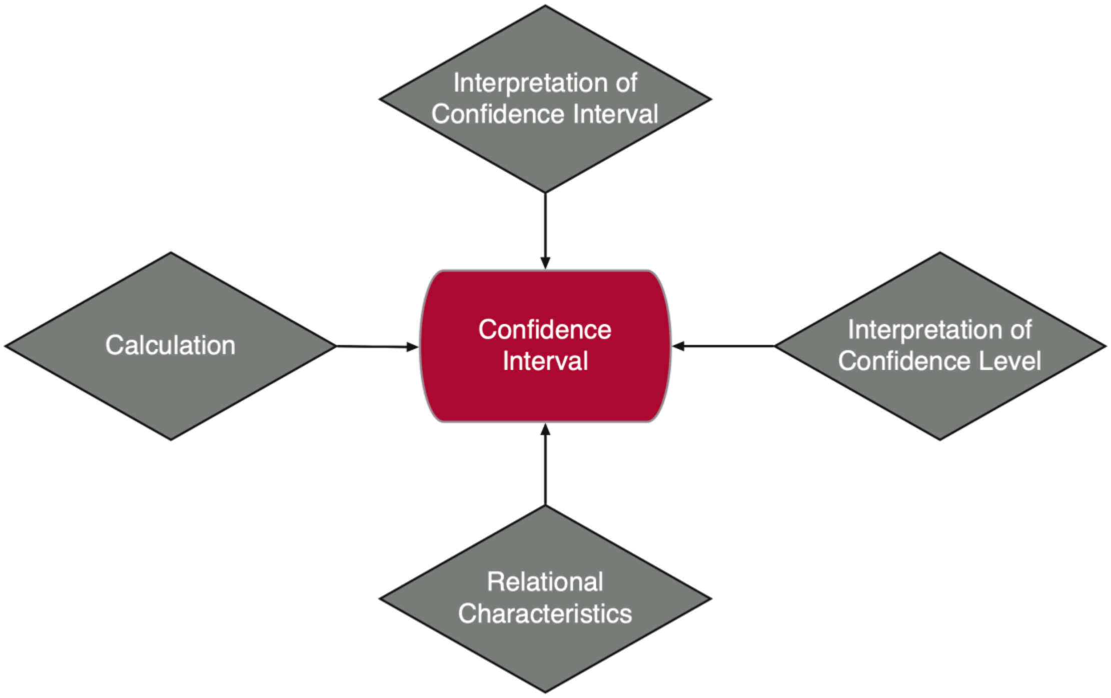
Concept Image and Developmental Cloud for Confidence Intervals

The theoretical underpinnings of a confidence interval require learners to simultaneously coordinate meanings and ways of thinking across multiple curricular concepts in statistics. In other words, learners need to be coordinating several learning clouds to develop a robust understanding of the concept of confidence intervals. From Section *Theory of Confidence Intervals*, there appeared to be four main components that incorporate understanding of confidence intervals: 1) calculation of a confidence interval, 2) interpretation of a confidence interval, 3) interpretation of a confidence level, and 4) the interpretation of relational characteristics. In the review of the literature, there are 3 major categories of (mis)conceptions about confidence intervals that have been studied: definitional, relational, and comparison. These (mis)conception studies provide little knowledge concerning the types of meanings and ways of thinking that are required for either deep understanding or misunderstanding of confidence intervals. This dissertation study aims to provide an initial concept image as a framework upon which future studies could be build. Due to the complex nature of the connections of learning clouds and the curricular concepts needed to deeply understand the concept of confidence intervals, it is difficult to conduct a conceptual analysis (see: Thompson, 2008). Instead, this section contains a general framework for an individual's concept image of the concept of confidence intervals. The need to describe the complex nature of learning clouds organizing the curricular concepts needed for a robust understanding of the concept of confidence intervals, the interpretation of confidence intervals, and the interpretation of confidence levels warrants the use of concept images and developmental clouds described in the previous section. Concept images and learning clouds provide the cognitive structure for constructing personal concept definitions of the concept of confidence interval, of the interpretation of a confidence interval, and of the interpretation of a confidence level. It is important to point out that this dissertation study does not intend to propose a learning progression for the conceptualization of confidence intervals, interpretation of confidence intervals, and interpretation of confidence levels. These proposed cognitive structures are intended to demonstrate the parallel and circular development of ideas that may be required to robustly understand the concept of confidence intervals, the interpretation of a confidence interval and the interpretation of a confidence level.

The first proposed concept image and developmental cloud (visualization of the concept image) for the concept of confidence intervals is decomposed into four main sub-concept images: 1) calculation of a confidence interval, 2) interpretation of a confidence interval, 3) interpretation of a confidence level, and 4) relational characteristics. The developmental cloud for the concept of confidence intervals is presented in Figure 3. The gray diamonds represent the four sub-concept images that are being proposed in this dissertation study. The GAISE: College Report 2016 recommended the move from computation heavy (procedural) instruction to instruction focused on the concepts of statistics (conceptual) (ASA GAISE College Report Revision Committee, 2017). Additionally, Steel, et al. (2019) encouraged instruction to be focused on what confidence intervals represented rather than the procedures used to create them. Therefore, the rest of the discussion of the concept images and developmental clouds for confidence intervals focuses on the concept images associated with the interpretations and practical implications of confidence intervals rather than the calculation learning cloud. Additionally, since the interpretation of the relational characteristics are not part of this dissertation, the sub-concept image and respective developmental cloud will not be addressed. This section begins with a brief description of the broad curricular concepts, based on a review of advanced and introductory statistics textbooks, that were needed to understand confidence intervals robustly and concludes with specific frameworks for each sub-concept image.

Figure 3

Developmental Cloud for the Concept of Confidence Intervals


Interpretation of Confidence Interval
Calculation
Confidence Interval
Interpretation of Confidence Level
Relational Characteristics

Hypothesized Concepts

This section contains proposed learning clouds constructed from my understanding of the pre-requisite and co-requisite knowledge needed for a robust understanding of the concept of confidence intervals. I propose five learning clouds that I have used to describe the coordination of the hypothesized pre-requisite and co-requisite knowledge: 1) parameter/statistic, 2) random process, 3) estimator/estimate, 4) sampling distribution, 5) coverage probability. Requisite curricular concepts for statistical inference start with understanding the underlying purpose of inference: to use information provided in a sample to understand a characteristic or attribute of a population better. The information typically derived from a sample is a statistic, as described in the previous section *Theory of Confidence Intervals*, and should be an unbiased estimator of the unknown value of a parameter. Thus, the first proposed learning cloud is *Parameter/Statistic*. Within this learning cloud, individuals need to understand the difference between the unknown value of the parameter, which is a characteristic of the population, and the actualized value of the statistic, which is a value calculated from the sample data. Before an individual can begin to understand the difference between the concepts of parameter and statistic, the individual must understand the difference between the concepts of a population and a sample. Furthermore, the individual needs to coordinate information about: 1) the population to which results can be generalized, 2) a sample, a variable produced from a sample, 3) a summary measure for the attribute of interest from a sample (statistic), and 4) a summary measure for the attribute of interest from a population (parameter). Therefore, the ideas that an individual must be coordinating are as follows:

- An *observational unit*: the person, animal, object, procedure, etc.
- *Datum*: an attribute of a unit of interest.
- *Variable*: a collection of data that could be either quantitative or categorical in nature.
    - *Quantitative variable*: a collection of numeric data on which arithmetic makes sense.
    - *Categorical variable*: a characteristic attribute.
- *Population*: the collection of all observational units of interest.
- *Parameter*: a numeric summary of the variable for a population.
- *Sample*: a subset of observational units taken from the population.

- *Statistic*: a numeric summary of the variable calculated from the data provided by a sample.

It is the interconnectedness of these concepts that produces the learning cloud of *Parameter/Statistic*. The (mis)conceptions concerning the interval as a range of individual scores and a range of values for the statistic may be caused by incomplete connections among these concepts.

The second learning cloud is *Random Process*, which encompasses the random process associated with a random variable. A random variable is produced through a random experiment designed to model a real-world situation. The model helps explain the randomness that exists in the population and sample data. As such, probability exists with respect to the model, rather than with the actual data. Understanding the difference in allocation of probability upon the random variable that is designed to model the real-world data versus the lack of probability on the real-world data is pivotal in understanding the idea of random process. This is particularly important when discussing the ideas surrounding inference: the statistical term for using a model based on data from a sample to make conclusions about a population of interest. The purpose of this learning cloud is to describe the ideas of probability and stochastic processes (models designed to explain a real-world situation) upon which most of the field of statistics is derived. The following curricular concepts may require simultaneous or prior development of the Parameter/Statistic learning cloud:

- *Random variable*: a variable that is produced from a random experiment.

    - *Random experiment*: an experiment or trial whose outcome is not predictable in the short term, but for which the long-run relative frequency of outcomes of different types in repeated trials is predictable.
        - *Outcome*: an unknown, but possible value (or attribute) of a random experiment.
        - *Relative frequency*: the number of data of one outcome divided by the total number of data for all possible outcomes.
- *Independent random variable*: a random variable that was produced from repeated independent trials of a random experiment.
    - *Independent*: the idea that the result(s) from a previous trial had no effect on the results from the current trial.
- *Probability*: the long-run relative frequency of a random variable.
- *Random sample*: a sample that was generated through a process that ensures, to the best of the researcher's ability, that every possible sample from the population had an equal probability of being selected.

This learning cloud is proposed to also be the foundation of the next learning cloud, Estimate/Estimator, because of the nature of the difference between an estimator and an estimate. The five conceptions of confidence intervals, 1) the proportion of the population, 2) the probability of the unknown value of a parameter within the interval, 3) the proportion of sample statistics, 4) the probability of statistic within the interval, and 5) the fixed interval, are hypothesized to occurred because of incomplete or incorrect knowledge within the concepts encompassed by the Random Process learning cloud.

Understanding the interpretation of an actualized confidence interval is reliant on understanding random variables versus fixed (but unknown) values. The next learning cloud *Estimator/Estimate* forms the basis of understanding the difference between random variables and fixed values. The most common (mis)conception of a confidence interval is the belief that the confidence level is the probability that the fixed, but unknown, actual value of the parameter is between the upper and lower bounds of the interval. This interpretation changes the random interval into a fixed interval and the fixed actual value of the parameter into a random value of the parameter. Both changes are incorrect. The endpoints of the interval are random values based on the random statistic, which is based on the random variable(s) that is used to model the real-word data. The actual value of the parameter is not known to the researcher, which explains the reason for inference, but the parameter of interest is defined. Understanding the distinction between the fixed unknown parameter and the random endpoints of the interval is not trivial and difficulties with this distinction may exist due to incorrect or incomplete knowledge of probability and random variables. Clarifying this conception might be difficult because prior conceptions can inhibit students from assimilating the new information concerning confidence intervals. This distinction in the definition of a confidence interval is difficult to communicate even for experts (Belia et al., 2005). The way that confidence intervals are written, $inf_{\theta} P_{\theta}(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$, implies that the value of the parameter is random, but it is important to recall what was random in this equation rather than relying on the written form, as Casella and Berger (2002) pointed out. Thus, the concepts being coordinated within this learning cloud are:

- *Random variable*: a variable that is produced from a random experiment.

- *Actualized variable*: a variable calculated from collected data.
- *Estimator*: a function of a random variable intended to be an approximation for the value of a parameter.
    - *Unbiased estimator*: an estimator for which the expected value (or mean) of the estimator was the value of the parameter for which it was intended to be an approximation.
    - *Point estimator*: a function of a random variable from a theoretical sample from a population.
    - *Interval estimator*: a random interval derived from two functions of a random variable from a theoretical sample from a population.
- *Estimate*: a function of an actualized variable that was intended to be an approximation for the unknown value of a parameter.
    - *Point estimate*: an actualized point estimator computed from an actual sample selected from a given population.
    - *Interval estimate*: an actualized interval estimator computed from an actual sample selected from a given population.

In addition to these learning cloud-specific concepts, an individual needs to coordinate the Parameter/Statistic and Random Process learning clouds. Little to no work had been done to explore the distinction between the estimator and the estimate – but it holds the potential to be a very important learning cloud for deep understanding of confidence intervals.

The *Sampling Distribution* learning cloud encompasses many concepts including the sampling distribution, which is part of the theoretical basis for confidence intervals.

The connection between confidence intervals and sampling distributions is very complex. The sampling distribution models the relationship between statistics from samples and the value of the parameter from the population. The behavior of the random statistics can be modeled by the sampling distribution, either theoretical-defined or simulated. This model forms the basis of the theoretical derivation of the formulas for confidence intervals. For an individual, the Sampling Distribution learning cloud may incorporate: 1) properties of distributions (measures of center, measures of variability, shape, probability); 2) differences between distributions of populations, distributions of sample, and distributions of statistics (sampling distributions); and 3) characteristics of sampling variability (i.e. between-sample variability). Thus, the concepts that are being coordinated within this learning cloud are:

- *Mean*: a measure of center for a quantitative variable.
- *Variability*: a measure of the how data values typically deviate from a center value
- *Standard deviation*: a measure of variability that was specifically the square root of the sum of the squared deviations.
    - *Deviation*: the difference each value of a variable was from the mean of the variable.
- *Distribution*: a representation of a variable.
    - *Theoretical probability distribution*: the formulaic representation of the probability of a random variable.

    - *Empirical probability distribution*: the actualized representation of the long-run relative frequency of a random variable based on a collection of outcomes from a random experiment.
    - *Shape of a distribution*: a description of the general behavior of the distribution (e.g. symmetric, skewed, uniform, unimodal, bimodal, etc.).
- *Distribution of a population*: the graphical representation of all values of the variable of interest from a population.
- *Distribution of a sample*: the graphical representation of all values of the variable of interest from a sample.
- *Sampling distribution of a statistic*: the theoretical distribution all possible estimates of a statistic calculated from all possible samples of size n.
    - *Sample size*, n: the number of units of interest in the sample.
    - *Mean of a sampling distribution*: the expected value of all possible estimates for a given statistic.
    - *Standard deviation of the sampling distribution for a statistic*: the theoretical standard deviation of all possible estimates for a given statistic.
    - *Standard error of the sampling distribution of a statistic*: an approximation of the standard deviation of the sampling distribution of a statistic based on an actualized statistic.

This learning cloud is also connected to the next learning cloud, Coverage Probability. The coverage probability is theoretically rooted in the sampling distribution. Limited knowledge concerning the sampling distribution, or the complexities of the sampling distribution, may be the root of documented incorrect conceptions such as: 1) thinking the confidence interval provides a range for the population data, 2) the relationship between sample size and confidence width, 3) the relationship between confidence level and confidence width, and 4) thinking the confidence interval contains the sample mean with the confidence level probability.

The final hypothesized learning cloud is the *Coverage Probability* learning cloud. This learning cloud is the foundation of the theory of a confidence interval. Within this learning cloud, the individual is hypothesized to be coordinating the ideas of Parameter/Statistic, Sampling Distribution, Random Process, and Estimator/Estimate but specifically within the context of a confidence interval. The Coverage Probability learning cloud is hypothesized to contain the ideas of confidence level, the confidence interval estimator, and the confidence interval estimate. Therefore, the specific concepts that an individual may be coordinating as part of this learning cloud are:

- *Confidence interval estimator*: a form of an interval estimator that had been derived based on the desired coverage probability.
    - *Coverage probability*: the probability that the interval estimator would capture the value of the unknown parameter from a given population.
    - *Confidence level*: the proportion of statistics from the middle of the true sampling distribution for the statistic generated from samples of size n centered at the actual value of the parameter, i.e. middle confidence level % of the sampling distribution.
- *Confidence interval estimate*: a form of an interval estimate, with a coinciding confidence level, derived from a confidence interval estimator.

Problems with the development of this learning cloud may be evident in all of the (mis)conceptions documented in Table 3. The next two sections focus on how these main learning clouds are connected within hypothesized concept images of the interpretation of a confidence interval, and the interpretation of a confidence level.

Concept Image and Developmental Cloud for Interpreting a Confidence Interval

The interpretation of a confidence interval can be written as: We are $100(1-\alpha)\%$ confident that the actual value of parameter is between (lower limit) and (upper limit). An individual needs to be coordinating several learning clouds in order to unpack what the meaning of this sentence is:

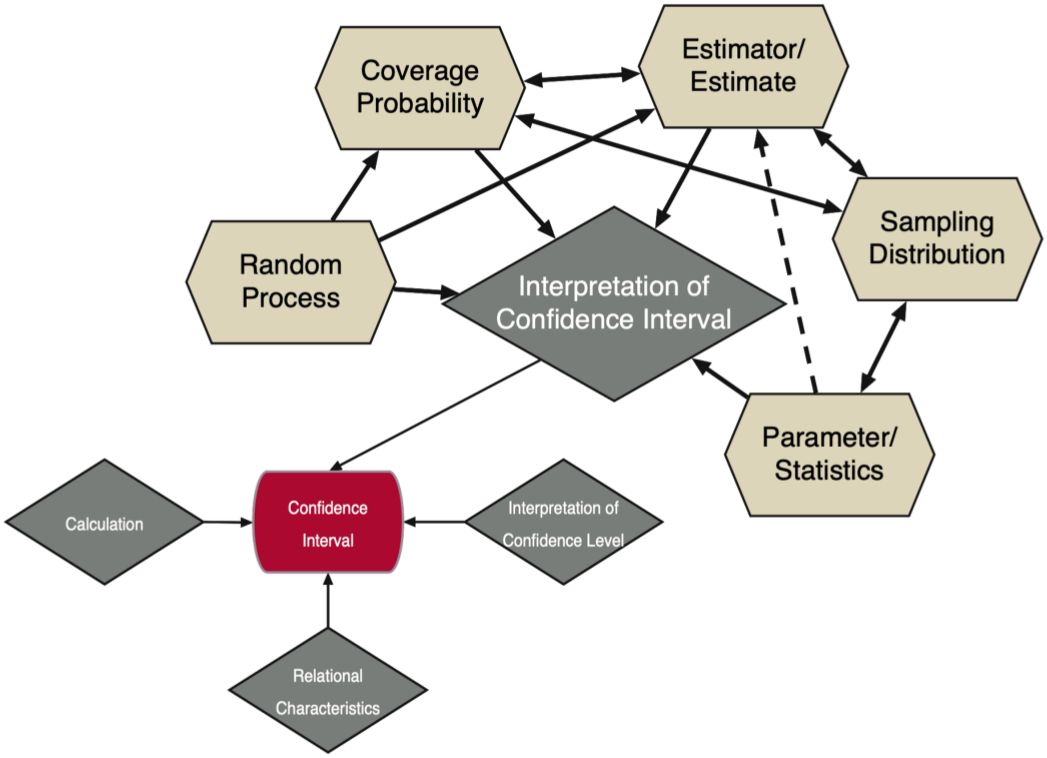
- $\mathbf{100(1-\boldsymbol{\alpha})\%}$ refers to the coverage probability that the interval estimator will capture the unknown value of the parameter of interest and requires the Coverage Probability learning cloud.
- **Confident** refers to the random process and coverage probability used to derive the confidence interval and requires the Random Process and Coverage Probability learning clouds.
- **Parameter**, which is usually stated in context, refers to a fixed, but unknown, value that summarizes a variable of interest (or attribute) within a population and is contained within the Parameter/Statistic learning cloud.
- **Between** refers to the fact that the actual value of the parameter is fixed and the the interval "captured" the unknown value of the parameter not that the unknown value of the parameter was moving (see for more information: Callaert, 2007; Foster, 2014). This requires understanding the Estimator/Estimate and Random Process learning clouds.

- **(Lower limit) and (Upper limit)** are possible values for the unknown value of the parameter which is part of the Parameter/Statistic and Estimator/Estimate learning clouds.

Figure 4 displays a possible developmental cloud for the Interpretation of a Confidence Interval (Confidence Interval Interpretation) concept image. This figure shows the interconnectedness of the learning clouds, displayed as tan hexagons, and how an individual may have to coordinate many different ideas to fully understand the interpretation of a confidence interval. The four main learning clouds connected to the interpretation of a confidence interval are: Parameter/Statistic, Estimator/Estimate, Coverage Probability, and Random Process. Underlying the Coverage Probability and Estimator/Estimate is the learning cloud Sampling Distribution.

Figure 4

Developmental Cloud for Confidence Interval Interpretation



For Figure 4 and 5, uni-directional arrows symbolize a one-way coordination of processes and bi-directional arrows symbolize a two-way coordination of processes. The coordination of processes is explained next. The relationship between the sampling distribution and the population distribution is related to the Parameter/Statistic learning cloud, as an individual must have a fairly developed Parameter/Statistic learning cloud in order to understand a sampling distribution. The ideas within Coverage Probability and Random Process must be coordinated with the Estimator/Estimate learning cloud, which is used to understand the difference between the words confident and probability in the interpretation of a confidence interval. Lastly, the Coverage Probability learning cloud incorporated many of the ideas of the other four learning cloud (Random Process, Estimator/Estimate, Parameter/Statistic, and Sampling Distribution). In this concept image, the Coverage Probability learning cloud was specifically used to identify the confidence level and what the confidence level meant. Together, these learning clouds can help produce a deeper understanding of the interpretation of a confidence interval.

Concept Image and Developmental Cloud for Interpreting a Confidence Level

This next section focuses on the interpretation of a confidence level. Traditionally, the interpretation is: "Approximately $100(1-\alpha)\%$ of all samples of size n from a given population are expected to produce $100(1-\alpha)\%$ confidence intervals that will contain the actual value of the parameter of interest." Like the interpretation of a confidence interval, this interpretation is hiding several concepts that need to be unpacked:

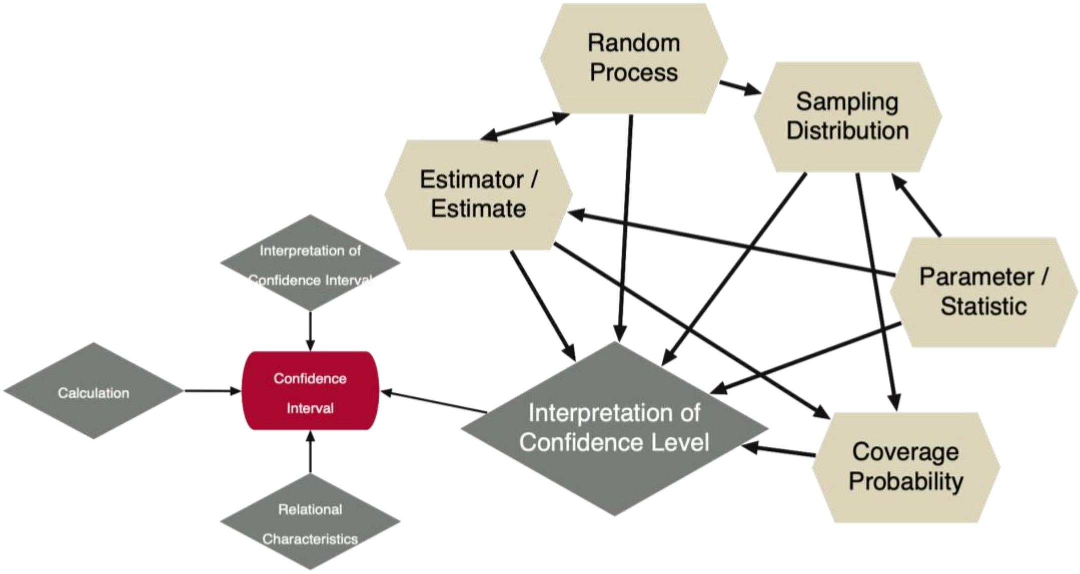
- **Approximately** refers to the long-run probability, the difference between theoretical and empirical distributions, and the difference between knowing the standard deviation of the sampling distribution or the standard error of the sampling distribution.

- $\boldsymbol{100(1-\alpha)\%}$ refers to the coverage probability of the given confidence interval.
- **All samples of size n from a given population** requires the learning clouds Parameter/Statistic, Sampling Distribution, and Random Process to understand the ideas of gathering all of the samples possible from a population.
- **Will produce** refers to the Random Process and Estimator/Estimate learning clouds which allows an individual to understand the difference between an estimator and an estimate.
- $\boldsymbol{100(1-\alpha)\%}$ **confidence intervals** refers to the specific confidence interval of a given confidence level that had been constructed from each sample. This requires the Random Process and Estimator/Estimate learning clouds.
- **Will contain** refers to the ideas that the actual value of the parameter is fixed, and the interval is random. This requires the Estimator/Estimate learning cloud.
- **Parameter of interest** requires the Parameter/Statistic learning cloud.

The final part of the sentence that requires unpacking is the “$100(1-\alpha)\%$ of all samples … $100(1-\alpha)\%$ confidence intervals.” These two percentages should to be the same for the sentence to hold true. It is still a true sentence if the first “$100(1-\alpha)\%$” was larger than the second “$100(1-\alpha)\%$”: “approximately 98% of all samples of size n from a given population will construct 95% confidence intervals that will contain the actual value of the parameter of interest.” In this case, we have overestimated the proportion of samples (98% instead of 95%) that should produce 95% confidence intervals that contain the actual value of the parameter of interest. This is not, however, the most productive interpretation of the confidence level. Therefore, the interpretation used in this dissertation study will assume equal proportions of confidence intervals that capture the actual value of the parameter and confidence level. Figure 5 demonstrates the interconnectedness of the learning clouds, tan hexagons, described above. When interpreting the confidence level, it is important to understand the ideas of the Coverage Probability and Estimator/Estimate learning clouds, which are the most prevalent learning clouds in this concept image. These two set the foundation for understanding what the confidence level is and how to practically communicate its meaning. Underlying these two main learning clouds were the ideas behind Random Process, Parameter/Statistic and Sampling Distribution learning clouds, which allows individuals to understand the repeated nature, stochastic process, and theory behind the coverage probability and interval estimator. These learning clouds could combine to help an individual develop a deeper understanding of the interpretation of a confidence level.

Figure 5

Developmental Cloud for the Confidence Level Interpretation



This chapter has provided hypothesized concept images for confidence intervals, the interpretations of confidence intervals, and the interpretation of the confidence level. It has also provided a summary of the current literature of the body of work surrounding confidence intervals, most of which were (mis)conception studies. These (mis)conception studies have helped motivate the need to understand the meanings and ways of thinking required for a deep understanding of confidence intervals, as well as the need to understand what meanings and ways of thinking have or have not formed that resulted in the documented (mis)conceptions of confidence intervals. I have identified four potentially necessary components to understanding confidence intervals: 1) calculation of confidence intervals, 2) interpretation of confidence intervals, 3) interpretation of confidence levels, and 4) relational characteristics. For this dissertation study, I generated hypothesized concept images for two of the four components: interpretations of confidence interval and confidence level. The next section will discuss the methods used to study these concept images.

CHAPTER 3

METHODS

In this chapter, I discuss the methodology and methods that are used to answer the research question posed for this dissertation study:

3. What conceptualizations of interpretations of confidence intervals and interpretations of confidence levels do undergraduate and graduate students have?
4. What similarities and differences exist in undergraduate and graduate students' concept images of the concept of confidence intervals when they conceptualize interpretations of confidence intervals and interpretations of confidence levels?

First, I discuss the research setting for this study, including a brief description of the courses from which participants were recruited, the recruitment methods, and the participants for this study. I then discuss the participant selection process. I conclude this section with a description of the methodology, data collection, and data analysis used in this dissertation study.

Research Setting

The data for this dissertation study were collected at a large research-focused institution in the southeastern part of the United States. The university, as of fall 2019, had approximately 39,000 students enrolled, with approximately 30,000 undergraduate students. The participants for this study were recruited from courses taught in the Department of Statistics and from currently enrolled graduate students. The Department of Statistics had undergraduate programs in statistics and data science and graduate masters' and Ph.D. programs in statistics. The department had three introductory statistics service courses (traditional, business, and life sciences). The department also had several service courses for graduate students. As a fairly large statistics department, with 32 permanent members of faculty, it offered over 30 different courses a semester, with many of the introductory courses having multiple sections.

The data collected for this study were gathered from students at three levels of statistics courses (introductory, intermediate, and senior) and statistics Ph.D graduate students. These data were collected from students who had previously completed the introductory statistics course in Summer 2019, students currently enrolled in the intermediate statistics course or the statistics major capstone course during Fall 2019, and students currently in the Ph.D program as of Fall 2019. The next section describes the courses and program from which the participants were recruited.

Introductory Statistics

The introductory statistics course was part of the general education requirements and was required by many majors on campus. It was a multi-section, four-credit course, which runs during the fall, spring, and summer sessions. It met for three hours in large sections (180 students). Typically, 3-6 instructors (faculty or graduate teaching assistants) were assigned to teach the 6 sections offered each semester. These lectures either consisted of three-50-minute lectures (Monday, Wednesday, and Friday), or two-75-minute lectures (Tuesday and Thursday). For the fourth hour, the students were divided into smaller classes of approximately 30 students and meet in a computer laboratory taught by graduate teaching assistants. During the time of the study, approximately 1200 students were enrolled in the six lecture and 36 laboratory sections. The laboratory sections used a manual, which I helped create and am a published second author on, designed to elicit conceptual connections among the statistical concepts covered during lecture. The course included topics such as: quantitative and categorical descriptive statistics, inference for one and two population proportions and means, inference for regression, and inference for two categorical variables. The assessments (homework and exams) for the course were completed through the online portal, WebAssign. The students were assessed through quizzes, weekly homework assignments, 3 within semester exams, and an optional final exam. Most assessment items were closed-form (fill in the blank, multiple choice, multiple select, true/false, select from the drop-down menu). StatCrunch software was used for statistical analysis and conceptual applets.

Intermediate Statistics

Intermediate statistics was a second course in the statistics department. It had an introductory statistics course as a pre-requisite, including AP credit for the AP Statistics exam. The course was an extension of an introductory course and included topics such as: multiple regression, analysis of variance, and non-parametric tests. It used JMP for data analysis and promoted communication and implementation of analysis using statistical software. As a three-credit course, it met for 2-75-minute sections on Tuesday and Thursday. There were typically three sections each semester, each section with a maximum of 40 students. Assignments included homework, quizzes, and mini projects. This course was a required course in the statistics major and minor. It was not required for students pursuing the data science major or minor. More advanced statistics courses, however, required this course as a pre-requisite.

Capstone Course

Statistics majors in the department were required to enroll in a year-long 3-credit capstone seminar that met in 3-50-minute classes (Monday, Wednesday, Friday) per week. The course centered around a project (a client-introduced problem, a self-selected problem, or a data repository problem), which was intended to last the entire year. Through these projects, students were encouraged to learn advanced statistical data analysis techniques, work with real data, and increase their communication skills. Students enrolled in this course were graduating students. The year-long course culminated in a poster session during which the groups presented their project findings.

Graduate Student Population

The graduate program in the statistics department had a masters and doctoral program. There were approximately 90 primary statistics masters and doctoral statistics students. Approximately two-thirds of the students enrolled in both programs were international students. The required coursework for both the masters and doctoral programs was strong in statistical theory, with less attention to applied coursework. The students in both programs were required to take a consulting course. Students also have the opportunity to work or volunteer for a department run consulting center.

Recruitment Methods

Study participants were recruited in one of two ways: 1) email sent via course instructor or coordinator or 2) in-person during the regularly scheduled class meeting. At the beginning of the fall 2019 semester (8/19/2019), I emailed course instructors for the introductory statistics, intermediate statistics, and capstone courses (see: Appendix B for the forwarded recruitment script) to ask for their help recruiting students from their classes (see Appendix C and Appendix D for the recruitment emails for introductory students and all other students, respectfully). For the introductory statistics recruitment, I asked the coordinator to email the Summer 2019 students. Intermediate and capstone students were currently enrolled in their respective courses in Fall 2019. The graduate students were recruited from the listserv current as of Fall 2019. A month after the initial email (9/16/2019), I emailed the course contacts (see Appendix E). Introductory, capstone, and graduate students were forwarded the recruitment emails in Appendix C and Appendix D. The course contact for the intermediate students requested an in-person recruitment rather than an email recruitment. On 10/16/19, I read the script provided in Appendix F to the three sections taught during fall 2019. All participants were asked to contact me via email or text. All but three participants arranged meetings through email, the remaining three texted. All interviews took place on campus in a building of mutual agreement.

Participants

Table 4 contains the number of study participants. There were originally twenty-one Interview 1 participants, seventeen Interview 2 participants, and sixteen Interview 3 participants. For the purpose of this study, participants were only considered for analysis if they completed all three interviews, resulting in 16 possible participants. From these participants, one introductory participant was removed because of audio malfunction during one of the interviews, one capstone participant self-removed from the study after the third interview, and one graduate participant was removed due to a language barrier. During initial review of the data, two more participants were removed: one introductory and one capstone student. The introductory student demonstrated very little recalled knowledge of confidence intervals, interpretations of confidence interval and confidence level. As the purpose of the study was to uncover the knowledge required to understand confidence intervals, basic knowledge of confidence intervals was required, which the introductory student did not meet. The capstone student was a dual major, statistics and a major of philosophical nature. The participant had a different perspective of probability from a traditional student, which would make this participant more of an anomaly than a typical student. The final breakdown of participants can be found in the final column in Table 4. Therefore, the transcripts of eleven participants were analyzed.

Table 4

List of Study Participants by Statistics Courses

Statistics Courses	Interview 1	Interview 2	Interview 3	Final
Introductory	5	3	3	1
Intermediate	5	3	3	3
Capstone	6	6	6	4
Graduate	5	5	4	3
Total	21	17	16	11

Characterizations of Participants

Of the original three introductory students, Tiana (all names are pseudonyms) was the only student analyzed for this dissertation study. Tiana was a pre-med student, majoring in biology. Her[3] perception of statistics was that of a simplistic, procedural nature. She appreciated the "plug and chug" nature of the course she had taken.

[3] Demographic information (age, gender, race, ethnicity, preferred pronouns etc.) were not collected over the course of the interviews. The use of gendered pronouns and pseudonyms should be understood to relate to the assumptions the author made of the participant's gendered identity. Thus, any reference to gender is part of my inferred persona of the participant, rather than to the individual's gendered identity.

Kiara, Aiden, and Logan were the three intermediate students. All three participants were part of the School of Business. Kiara and Aiden were anticipating graduating May 2020, while Logan was anticipating a May 2021 graduation. Kiara was applying for graduate programs in the School of Business. Aiden and Logan were both uncertain of their future plans but had both recently registered for the new data science minor through the Department of Statistics. All three intermediate students had taken two other statistics courses in the School of Business. As such, each had had at least three statistics courses, including AP Statistics, prior to the intermediate course they were enrolled in at the time of their participation in the study. These three participants had a business analytics view of statistics: hypothesis testing, modeling, and procedural view of statistics. At the time of the first interview, Kiara admitted to a year and a half hiatus from statistics curriculum. Aiden and Logan both held strong hypothesis testing conceptions through the interviews, regardless of the context.

The capstone students were Diana, Brody, Emma, and Gabe. As graduating seniors in statistics, they had completed most of the statistics course work. Gabe, Diana, and Brody had completed both mathematical statistics courses by the time they participated in the study. Emma and Brody had both tutored student athletes in introductory statistics. Emma and Diana were dual majoring in statistics and in a major in the School of Business. Brody was also dual majoring: statistics, and in a program of philosophical nature.

Liam, Jace, and Joel were graduate students currently enrolled in the statistics PhD program. Liam had a previous MS in Actuarial Science, a BS in statistics, and was a third-year student in the PhD program. Liam had completed the first two years of core coursework and was beginning to work on his elective coursework. Additionally, Liam had been assigned as a teaching assistant for the laboratory sessions associated with the introductory statistics course. Jace had a prior engineering degree and decided to change careers to statistics. He started at his current university in the statistics MS program, had transferred to the statistics PhD program and was working on his second-year PhD coursework at the time of the study. Joel was also a third-year student and had a BS and MS degree in statistics. Joel was also working on finishing his elective coursework. The diverse backgrounds of the participants provided me with different views and knowledge concerning conceptualizations of confidence intervals, interpretations of confidence intervals, and interpretations of confidence levels.

Methodology

This next section discusses the methodology used to design the tasks and the research methodology used in this study. This dissertation study was designed to elicit participant knowledge of confidence intervals. In particular, the study focused on participants' conceptualizations of confidence intervals. The primary method of data collection was through clinical intervals. The first subsection discusses the design and implementation of the clinical interviews. The second subsection discusses the specific tasks used in this study.

Task Design

The primary method used to answer the guiding research question for this dissertation study was through clinical interviews (e.g. Goldin, 2000; Hunting, 1997) spread over three one-hour sessions. The interviews were a structured task-based interview as described by Goldin to develop a better understanding of the concepts and connections participants already held. Using a semi-structured interview protocol ensured that all participants would answer the same general questions but allow me the freedom to deviate from the protocol should follow up questions be required. By interviewing the same participant over multiple time points, I was able to track the participant's thinking, connections, and concepts across multiple sessions. This also allowed me to gain the participants' trust in order to have better interactions (Clement, 2000; diSessa, 2007). All of the questions in the protocols would be considered open-middle tasks (Bell & Burkhardt, 2002; Yeo, 2017) as they all had correct answers. The questions were mostly conversational in nature, allowing the participant to take whatever path came to mind to answer the question. The questions were designed to elicit the conceptualizations each participant had concerning the interpretation of confidence interval, interpretation of the confidence level, and relational characteristics.

Task Descriptions

I designed three protocols for this dissertation study. The first two interviews were designed to elicit student knowledge concerning the conceptualizations that students had about confidence intervals. The third interview consisted of a task designed to introduce students to confidence intervals in an introductory statistics course.

The first interview, see Appendix H, was designed to elicit student knowledge about the statistical concepts in their conceptualization of confidence intervals. These questions were designed to elicit connections students had formed about their knowledge of inference, construction of a confidence interval, interpretation of confidence intervals, interpretation of confidence levels, and characteristics of confidence intervals (see

Appendix H for rationale). These questions were open-middle questions and were mainly conversational in implementation.

The second interview, see Appendix I, focused on interpretations of confidence intervals, interpretations of confidence levels, differences between estimators/estimates, explorations of the meaning of the confidence level, and practical understandings of relational characteristics. The interpretation questions were in the form of a series of sentences. Each slide had four to five sentences that participants were asked to explain. Most of the sentences exhibited known (mis)conceptions of the interpretations of confidence intervals and confidence levels. Four web-based applets were used to explore the meaning of the confidence level (see: Appendix I). The purpose of the sentences and applets was to provide participants with visuals and non-normative statements to expose any missing connections or understandings. The rationale for each task can be found in Appendix I.

The final interview, Interview 3, can be found in Appendix J. The purpose of this interview was to implement a task designed to introduce introductory students to confidence intervals based on the missing concepts and connections that were seen through the first two interviews. All participants were given this task. Advanced students were told this was a task designed to introduce confidence intervals to introductory students and asked to engage as authentically in the task as possible. This task connected the ideas of coverage probability to the margin of error (and corresponding confidence coefficient). It was intended to help students identify the difference between the random aspect of the estimator and the fixed aspect of the estimate, which should help students understand the difference between the words *probability* and *confident* in the interpretation of a confidence interval. Lastly, it was designed to help students visualize the ideas of a confidence interval and provide a more theoretically sound introduction than is traditionally taught in introductory statistics courses.

Data Collection

The interviews for this study were conducted during the fall 2019 semester. Participants were contacted at the beginning of the semester (August 2019). First interviews began August 23, 2019. Due to staggered recruitment, first interviews were conducted until November 5, 2019. Second interviews began October 21, 2019 and concluded on November 15, 2019. For most participants, there was approximately a one-month gap between the first and second interview. Late recruits experienced a one- to two-week gap between interviews. Third interviews began November 8, 2019 and concluded on December 13, 2019. Most participants experienced a two- to three-week gap between second and third interviews. Appendix K provides the dates of the interviews with the participants of the study.

The data collected during the task-based clinical interviews consisted of two video recordings and up to two screen recordings, including audio. One participant was recorded on an additional audio recording device due to the participant's soft-spoken nature and difficulties in audio recordings on Interview 1. Participants worked on my iPad Pro, which was connected to my computer, an Apple Macbook Pro, and was screen-recorded using QuickTime. The participant's raw work for the three interviews was captured in an app for the iPad, Goodnotes4. A second Apple Macbook Pro was used for Interview 2 to record the screen actions during the applet use required for the interview protocol. The screen was recorded using QuickTime. FinalCut Pro was used to create video files for analysis. Audio was saved separately and uploaded to Otter.ai (https://otter.ai) for audio transcription. MaxQDA was used to analyze the data.

The entire data set consists of approximately 28 hours of interviews. Table 5 provides the detailed list of length of interviews. For the purpose of this study, Interviews 1 and 2 were analyzed. The protocol for Interview 3's did not address the research questions of this study directly. If something arose in the first two interviews that could be further explored by Task 1 of the third interview, these segments were included in the analysis. Thus, the final analyzed data set consisted of approximately 20 hours of interviews.

Table 5

Length of Interviews

Participant	Statistics Course	Interview 1	Interview 2	Interview 3	Total
Joel	Graduate	56:38	58:54	49:22	2:44:54
Liam	Graduate	57:16	51:57	38:15	2:27:28
Kiara	Intermediate	39:45	57:14	42:49	2:19:48
Aiden	Intermediate	49:12	42:32	34:54	2:06:38
Diana	Capstone	1:04:27	1:05:22	55:01	3:04:50
Emma	Capstone	45:37	54:16	36:59	2:12:52
Mia*	Introductory	41:22	54:34	1:04:20	2:40:16
Jace	Graduate	43:49	58:30	1:01:35	2:43:54
Tiana	Introductory	52:37	51:43	1:03:16	2:47:36
Mason*	Capstone	35:42	38:01	43:16	1:56:59
Brody	Capstone	46:55	47:54	29:09	2:03:58
Logan	Intermediate	36:15	54:54	50:38	2:21:47
Gabe	Capstone	35:49	50:08	36:15	2:02:12
Total		10:05:24	11:25:59	10:05:49	31:37:12
Final 11 Total		8:48:20	9:53:24	N/A	19:41:44

Note. * indicates removed from study after initial analysis.

Data Analysis

There are many possible ways to analyze video and audio recordings of task-based clinical interviews. These analysis methods include verbatim transcription, line-by-line analysis, and moderate analysis of video recordings (Koklu, 2017). Little was known about the connections students make when conceptualizing and reasoning about confidence intervals allowing for the use of a *generative* research approach, defined by Clement as "generat[ing] new observation categories and new elements of a theoretical model in the form of descriptions of mental structures or processes that can explain the data" (2000, pp. 332–333). This study mimicked an emergent research design, similar to methods used in grounded theory. Instead of grounded theory-inspired methods, the methods proposed by Powell et al. (2003) were used as a guide to analyze the task-based clinical interview video recordings.

Following a review of literature, Powell et al. (2003) concluded that within mathematics education research, there was little discussion concerning how researchers analyze video recordings or methodological issues with video recordings. To counteract this lack of literature, the Robert B. Davis Institute for Learning (RBDIL) at Rutgers University published an analytical approach that was developed to aid their researchers when analyzing clinical interviews of individual learners or working groups thinking and reasoning about mathematical concepts (Powell et al., 2003). For this method of analysis, the authors chose an ethnographic lens for the analysis, using Erickson's (1992) statement:

> "when … events are rare or fleeting in duration or when the distinctive shape and character of … events unfolds moment by moment, during which it is important

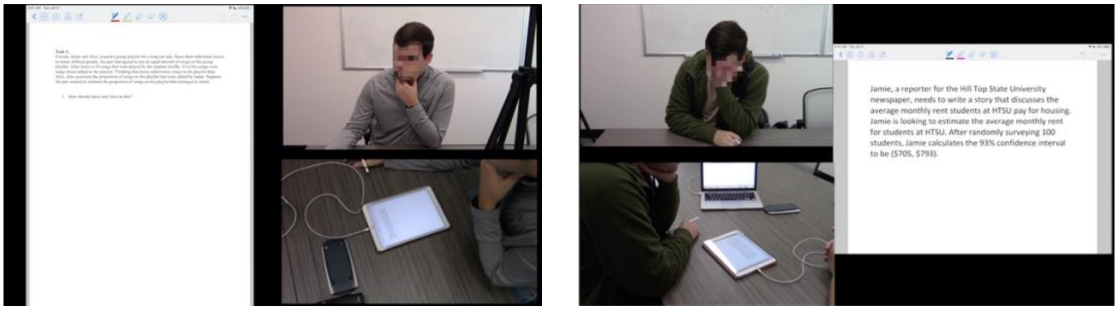
> to have accurate information on the speech and nonverbal behavior of particular participants in the scene … when one wishes to identify subtle nuances of meaning that occur in speech and nonverbal action—subtleties that may be shifting over the course of activity that takes place. (pp. 204–205)" (as quoted in: Powell et al., 2003, p. 413).

This lens has helped the researchers at RBDIL develop a non-linear pathway to analyze video data: 1) viewing the video attentively, 2) describing the video data, 3) identifying critical events, 4) transcribing, 5) coding, 6) constructing storyline, and 7) composing narrative (Powell et al., 2003, p. 413). It is important to note that these steps are meant to aid researchers in identifying phases of analysis rather than provide a step-by-step analysis process because of the sometimes cyclic and parallel nature of the steps in analysis (Powell et al., 2003). The analytical path through Powell et al.'s seven steps used in this dissertation study, along with detailed explanations of each of the steps, will be discussed in the paragraphs that follow.

The first step of the data analysis was to combine the videos acquired by three recording devices into a single video comprised of the three shots (Figure 6), using FinalCut Pro. The videos and the audio-only files were exported and saved separately. All audio-only files for the three interviews were edited and uploaded to Otter.ai. Using Otter.ai, edits were made to the transcripts from Interview 1 and 2. Transcripts with timestamps from Interview 1 and 2 were exported and uploaded to MaxQDA, along with the corresponding videos. This first step was a modification of the Powell et al. framework. *Transcribing* is intended to be solely done on *events*, which are identified as "connected sequences of utterances and actions that, within the context of our *a priori* or *a posteriori* research questions, require[d] explanation by us, by the learners, or by both" (Powell et al., 2003, p. 416, emphasis in original). Powell et al. further specify that a *critical event*, which is contextual with its relationship to the research questions, is when "[the event] demonstrates a significant or contrasting change from previous understanding, a conceptual leap from earlier understanding" (Powell et al., 2003, p. 416). Due to the complex nature of language in describing statistical ideas, identifying critical events would require the use of transcripts to help me follow the statements made by the participants.

Figure 6

Screenshots of Videos Used for Analysis



Task 1:

[illegible]

1. [illegible]

Jamie, a reporter for the Hill Top State University newspaper, needs to write a story that discusses the average monthly rent students at HTSU pay for housing. Jamie is looking to estimate the average monthly rent for students at HTSU. After randomly surveying 100 students, Jamie calculates the 93% confidence interval to be ($705, $793).

Once the video and transcripts were uploaded into MaxQDA, I proceeded with the next two steps of Powell et al.'s methodology, *viewing the video attentively* and *identifying critical events*. Here, the authors suggested that researchers watch, but not analyze, the videos. This allowed me to better attend to the data as a whole, seeing the whole picture, rather than attending to specific items within the data (Powell et al., 2003). During the initial full viewing of the videos, I coded of questions by question themes and poignant parts of the transcript by theme. Powell et al. suggested that *coding* is contextual based on the researchers' research questions. During this process, 11 question themes were coded, as shown in Table 6. Segments of text that corresponded to important learning clouds identified in the conceptual analysis discussed in Chapter 2 Concept Image and Developmental Cloud for Confidence Intervals were also coded: 1) Coverage Probability, 2) Sampling Distribution, 3) Random Process, 4) Estimator/Estimate, and 5) Parameter/Statistic. Finally, I coded any other seemingly significant segments of transcript based on knowledge of the existing literature.

Table 6

Coding of Question Themes and Learning Clouds

Memo Theme	Code	Question(s)
Interpret CI	Interpret CI	Interview 1: Question 3 Interview 2: Task 1: Question 1 Interview 2: Task 1: Question 3
	Plausible Range for the Value of the Parameter	Interview 1: Question 5
	Interpret Interval	Interview 2: Optional Task: Question 1
	How to Calculate	Interview 1: Question 2
Interpret CL	Interpret CL	Interview 1: Question 4 Interview 2: Task 1: Question 2 Interview 2: Task 1: Question 3
	Jamie's Colleague	Interview 2: Task 2
	Process to Generate	Interview 2: Task 1: Question 4
	Applets	Interview 2: Task 3
All	Parameter/Statistic	
All	Random Process	
All	Estimator/Estimate	
All	Sampling Distribution	
All	Coverage Probability	
None	How to Estimate	Interview 1: Question 1
None	Change in Sample Size	Interview 1: Question 6
None	Change in Confidence Level	Interview 1: Question 7

After the initial coding of all 13 participants, two participants (Mia and Mason) were removed from further analysis as described in the *Participants* section above. Of the remaining 11 participants, two groups were created: Group A consisted of five participants (Kiara, Tiana, Diana, Brody, and Joel) and Group B consisted of six participants (Liam, Aiden, Emma, Jace, Logan, and Gabe). The two groups consisted of participants that were recruited from different levels of statistics courses. Group A was selected to be analyzed further to develop an initial idea of a general concept image. The participants in this group were selected based on a pilot analysis done prior to the analysis for this study. A participant from each statistics course level that demonstrated different conceptualizations on The Jamie's Colleague's Task (see: Interview 2, Task 2 in Appendix I) was selected for pilot analysis. Brody was included in this group because his demonstrated knowledge was vastly different from the other four participants. These participants demonstrated different perspectives of knowledge of confidence intervals but were not assumed to be demonstrating vastly different perspectives from Group B.

With Group A, the next step in the analysis consisted of *describing the video data*, which was meant to help the researcher summarize the video recordings in a descriptive but not analytical way (Powell et al., 2003). The authors suggest that these be time-coded and refer to factual descriptions only ("he writes…") rather than inferential descriptions ("he is trying to…") (Powell et al., 2003, p. 416). Here, using MaxQDA's function to select documents and particular codes, summaries were written for each participant based on the research questions: interpretation of confidence intervals, and interpretations of confidence levels. These memos consisted of summary comments about the participants' statements, including documenting demonstrated knowledge and work by the participants. For each interpretation memo, question themes associated with the memo and all five of the important learning cloud codes (Table 6), were activated so that only these coded segments were visible for the summary. As these two memos were developed, it became evident that a separate memo needed to be developed titled: Estimator/Estimate. These memos were written for each participant in the following order: Brody, Kiara, Diana, Tiana, and Joel. At this stage, each participant had the following memos created: (Participant Name)– CI Interpretation, (Participant Name, i.e. Brody) – CL Interpretation, and (Participant Name) – Estimator / Estimate.

Following the initial summary of Group A, the next stage of analysis began: *constructing a storyline*. According to Powell et al. (2003), this allows the researcher to make sense of the data while attending to the coding. Constructing a storyline allowed me to begin to make connections and interpret the events that had been coded to provide meaningful insights into the participant's concept images. For the Group A participants, a fourth memo (titled: (Participant Name) - Summary) was created that analyzed each participant's conceptualizations of the interpretation of a confidence interval and interpretation of the confidence level across the two interviews. It also became evident through this process that a broader look into each participant's concept image of confidence intervals was needed. This memo included a section for "general thoughts" where I recorded these summary notes. This memo also contained links to relevant coded segments of transcripts. From these more analytical summary memos (title: (Participant Name) - Summary), a final collective memo was written, titled: General Thoughts. Here, the conceptualization of each participant's interpretation of a confidence interval and interpretation of a confidence level was summarized with a focus on the apparent similarities and differences in concept images from the other participants in Group A. At this point, it began to become evident that there were themes to conceptions of different aspects of the interpretations of confidence intervals and confidence levels. Separately, these themes (Confidence, Confidence Level, Probability/Random Process, Visualization, and Accuracy) were analytically summarized. To supplement the summarized memos (title: (Participant Name) - Summary), each participant's transcript, with a particular focus on any previously coded segments of transcript, were read again to add context and depth of analysis of the demonstrated knowledge.

Based on the initial analysis of Group A, the data from the Group B participants were analyzed in a less iterative process. These six participants (Liam, Aiden, Emma, Jace, Logan, and Gabe) were analyzed using the initial framework found through the analysis of the Group A participants. Separate memos (interpreting confidence intervals, interpreting confidence levels, and estimator/estimate) for each of Group B's participants were created simultaneously. Unlike the Group A's summaries, these focused on the themes that were found to be important in the initial round of analysis. Any new themes were noted. Immediately following the creation of these three summary memos (CI Interpretation, CL Interpretation, and Estimator/Estimate), the general analytical summary (Summary) was written, and each participant's conceptualizations were added to the general analysis memo (General Thoughts).

Once the General Thoughts memo contained all participants, I began to categorize the conceptions of the interpretation of confidence interval, interpretation of confidence level, and the emergent themes (confidence, confidence level, probability/random process, visualization, and accuracy). The linked segments of coded transcript that were attached to the Summary memo for each participant were deemed a new layer of *critical events*. These segments were used as evidence for the appropriate categorizations of the participants. The results of these categorizations are presented in the next chapter.

Chapter 4 is the final stage of the analysis model consisting of the *construction of a narrative report*, which, as Powell et al. (2003) discussed, is a never-ending stage from the inception of research questions to final write up. Formally, this final stage of analysis is typically seen as the written conclusion presented as a summary of findings. The writing of the narrative report was two-fold: The first consisted of writing a detailed summary of each participants' conceptualizations, which can be found in Appendix L. The second consisted of the story discussed within Chapter 4 about the interpretations of confidence interval, interpretation of confidence level, and relevant themes across participants' concept images.

CHAPTER 4

RESULTS

This section contains the results of this study. In the first section, categorizations of the conceptualizations of the interpretation of the confidence interval and the interpretations of the confidence level are discussed, addressing the first research question. The second section addresses the second research question containing a discussion of two of five aspects of participants' concept images that were sources of similarities and differences among the participants' conceptualizations of the interpretations of the confidence interval and confidence level: the word confident, and the idea of confidence level. The last section of this chapter contains a classification of the concept of confidence interval. A summary of each participant's conceptualization of the interpretation of the confidence interval, interpretation of confidence level, and relevant dimensions of concept images can be found in Appendix L.

Conceptualizations of the Interpretations (R1)

In this section, I will discuss statements made by the participants when they were asked to discuss interpretations of the confidence interval and interpretations of the confidence level. Interview 1, Questions 3 and 4 (see: Appendix H) were designed to elicit participants' knowledge of the interpretation of confidence intervals and the interpretation of confidence levels, respectively. During Interview 2, participants were given a series of statements about the interpretations of confidence intervals and interpretation of confidence level (see Task 1: Appendix I). Interview 2, Task 1, was designed to elicit participants' comprehension of the statements, many of which were written based on documented (mis)conceptions, and to provide talking points around which participants' understanding could be probed. Over the course of the two interviews, participants had the opportunity to discuss the ideas surrounding the interpretations in many different ways. Participants were categorized based on statements made over the course of their interviews. It became apparent that many participants held multiple conceptions of the interpretations of confidence interval and interpretation of confidence level – often at the same time and in conflict with other conceptions. These conceptions changed based on the context and question that was asked. Tall and Vinner (1981) identified these scenarios as different evoked concept images. Thus, the categorizations that follow are not mutually-exclusive categories.

Interpretation of Confidence Interval

During Interview 1, Question 3, which was worded as "How should Jamie and Alex interpret the 95% confidence interval you just calculated?", produced several types of responses from the participants (for full statements, see: Appendix L). The responses provided were categorized into four rough groups: correct, capture/not capture, long-run interpretation and probability, and confounding. These categories are themes that emerged based on common topics included in the statements about the interpretations of the confidence interval provided on both interviews. Each theme is summarized in this section with a description, exemplar statements, list of participants, and classification of productive and non-productive reasoning, when applicable. The categorization of participants into themes is summarized in Table 7.

Correct Interpretation of a Confidence Interval

Emma, Liam, Brody, Aiden, and Gabe interpreted the confidence interval using sentences that were typically considered correct: “We are confidence-level% confident that the actual value of the parameter is between the upper and lower bound of the confidence interval.” Emma’s interpretation was an exemplar of this category: “They [meaning Jamie and Alex] could be 95% confident that the true proportion of songs that Jamie loaded on that playlist is between 52.8% [and] 79.1% …”

All of the participants in this category included other explanations in their interpretations of confidence intervals, creating two additional sub-themes: hypothesis test conclusions and formulaic. The first sub-theme contains responses that make claims about the true proportion of Jamie’s songs on the group playlist. The context for Interview 1 suggested that perhaps Jamie and Alex had not put an equal proportion of songs on the group playlist. Emma and Gabe both commented on the hypothesis test context in their interpretation. They concluded their interpretations with statements about the fact that p=0.50 was not within the calculated interval, indicating that Jamie perhaps added more songs to the playlist, as demonstrated by the continuation of Emma’s statement above “… which we can see that 50%, meaning half of the songs, it's not within the interval, which gives us reason to believe that he [meaning Jamie] actually did load more than half of the songs.”

The second sub-theme contained responses that described the correct interpretation as “formulaic” (Brody’s explanation). Brody and Liam both discussed the statement used to interpret the confidence interval as a “fill-in the blank” type of sentence that can be modified to fit any context. Liam, a teaching assistant for the introductory statistics course, explained his written statement as if he were teaching it to a student:

> So, you have to discuss what the interval is capturing. So, you want to capture the true proportion of songs added by Jamie to the playlist. And since we didn't include the entire distribution in our creation of the confidence interval, then we can't be exactly certain that we capture the true proportion. So that's why we have to add this qualifier that we're 95% confident, and our confidence interval that we created, and then you have to say what is the interval, that is between .52 and .79.

Table 7

Categorization of Themes in Responses for Interpretations of Confidence Interval and Confidence Level

		Intro	Intermediate			Capstone				Graduate			Total
		Tiana	Kiara	Aiden	Logan	Diana	Emma	Brody	Gabe	Joel	Liam	Jace	
Confidence Interval	Correct			Yes			Yes	Yes	Yes		Yes		5
	Capture/Not Capture					Yes	Yes	Yes	Yes		Yes	Yes	6
	Long-run/ Probability			Yes	Yes	Yes				Yes		Yes	5
	Confounding	Yes	Yes										2
Confidence Level	Long-run	Yes				Yes	Yes	Yes			Yes	Yes	6
	Coverage Probability		Yes		Yes		Yes	Yes		Yes	Yes	Yes	7
	Percentage	Yes	Yes	Yes	Yes		Yes		Yes	Yes		Yes	8

Capture/Did Not Capture

Liam, Brody, Emma, Diana, Gabe and Jace all discussed the idea that once the actualized interval has been calculated, the confidence interval has either captured the actual value of the parameter or it has not. If an actualized interval captures the actual value of the parameter, the probability the interval captured the actual value of the parameter is 1. If the actualized interval does not capture the actual value of the parameter, the probability the interval captured the actual value of the parameter is now 0. These four participants used the capture/not capture rationale to explain why the interpretation of the [actualized][4] interval does not include a probability statement. Participants stated that there is not a confidence level probability the actual value of the parameter is within the interval because "it's not a 95% probability that it's going to be in there, because either it's in there or it's not" (Diana). This interpretation presented itself often during the first interview and inspired the task named Jamie's Colleague (Interview 2, Task 2). It was not clear that participants understood the intended meaning behind this explanation. The capture/not capture explanation was used as a fact, rather than demonstrating deeper connections to the difference between an estimator and an estimate. This was evident from Diana's further explanation of the capture/not capture statements. She stated confusion about the placement of probabilities on unobserved things leaving the impression Diana, and others, had simply heard this explanation without internalizing its meaning:

[4] I introduced the term actualized interval in Chapter 2 as a way to distinguish between the interval estimator and the interval estimate. Most participants did not make this distinction. The notation "[actualized]" is meant to help the reader understand the difference without implying the participant recognized the difference between an estimator and an estimate.

> It [meaning confident] doesn't work the same way probabilities work. Because, like, I think if I remember the reason correctly, it's just either that parameter like either the true value is in that interval, or it's not. … there's a probability of one or zero. … you can only have probabilities with unobserved things.

The Jamie's Colleague task pushed participants to explore the difference between the interval estimator and the interval estimate. Specifically, the colleague made the claim that there was a confidence-level probability that the interval estimator would capture the actual value of the parameter (i.e. prior to collecting data, there is a confidence-level probability the interval will capture the actual value of the parameter). The colleague continued by stating that after the data had been collected, the probability the interval captured the actual value of the parameter was either 0 or 1. Of the eleven analyzed participants, only five participants agreed with both statements. The results from this question are beyond the scope of this dissertation study and will be addressed briefly in the *Implications and Further Directions* section.

Long-Run Interpretation and Probability

The ideas of the long-run interpretation and probability associated with the confidence level were found in at least one of the responses of the following participants: Aiden, Logan, Jace, and Diana. Aiden and Logan both discussed the interpretation of the confidence interval using the idea that there existed *probability* in the interpretation: 95% chance or 95% sure the actual value of the parameter is between the upper and lower bound of the confidence interval. Aiden, however, ended his discussion of the interpretation of the confidence interval by stating: "I'm pretty sure it's [referring to chance] something you're never supposed to say." He concluded his interpretation by saying that he thinks the statement is supposed to include the word confident, but stated that the word confident seemed like an “arbitrary declaration,” suggesting a probabilistic interpretation. Jace, Joel and Diana provided the *long-run interpretation* of the confidence level when asked to interpret the confidence interval. Jace and Joel discussed the interpretation as a replicate of experiments, with Jace stating “that 95% confidence interval means like, you do these experiment (sic), like thousand times or 100 times, there are 95% chance that the proportion of the confidence interval contains the true parameters.” Diana, on the other hand, stated that the confidence level references the chance the interval would capture the actual value of the parameter. She was clear that the confidence level is not a probability:

> Uh, I believe the way I was taught is 95% of the time it will capture the true value. Not that there's a 95% probab..., like not that it's a probability of it being in there, but that if you ran you know, the simulation ever, hundred times 95 of those would have the true population parameter.

Confounding Ideas

Kiara and Tiana interpreted the confidence interval using several confounding ideas that were mixed together to produce an interpretation. This category is different from a nonsense category in that the responses presented here can be grounded in other concepts and instructional approaches that appear to be confounded within each participant’s understanding. Tiana confounded the statements about the interpretation of a confidence interval and the long-run interpretation of the confidence level, writing the following: “We are 95% confident that, if we took 100 sample tests, the true proportion would fall between this range.” During the first interview, Kiara had difficulties using statistical terms such as, population, sample, statistic, parameter. This was particularly evident in her interpretation of the confidence interval. Here, she confounded the ideas of statistics, sample values, population, parameter, and population values. It was unclear from this interpretation what Kiara really thinks the interval represents:

> 95% of your confident that 95% of the actual, I guess parameters, the values that you pull from the population, not the sample, your confidence that 95% of the parameter, or the actual values from the population that you're pulling from the populations fall within the interval that we hypothetically put together.

Statements from Interview 2

Based on the themes presented above, it might appear that not all participants would select the same interpretations of confidence interval statements as correct or incorrect on Interview 2, Task 1. As shown in Table 8, this was not the case. In general, all participants, except Jace[5], agreed with both of the statements: "Jamie is 93% confident that the actual mean monthly rent for all students at HTSU is between $705 and $793" and "Jamie is 93% confident that the actual mean monthly rent for all students at HTSU falls within $705 and $793." These statements were exactly the same except for the placement of the action on the parameter of interest. In the first statement, the wording implied that the actual value of the parameter is a fixed value that the interval has 'captured.' The second statement implied action on the part of the actual value of the

[5] Jace had difficulties with the words *capture*, *falls within*, and *between*. In the end, Jace did not agree with any of the interpretations of confidence interval given during Interview 2. I did try to get Jace to explain his understanding of the words, and his issues with using them as interpretations. Upon the conclusion of this analysis, I am still unclear about the interpretations Jace had of the words and statements provided in Interview 2. It may, ultimately, be an English to Chinese translation issue. Jace never provided an interpretation of the actualized interval throughout the two interviews.

parameter, stating the actual value of the parameter “falls within” the interval. The typical interpretation of the confidence interval requires that the statement imply that the interval is the random item that is ‘capturing’ a fixed actual value of the parameter (similar to throwing horseshoes, where the horseshoe is the interval and the picket is the actual value of the parameter).

Additionally, none of the participants agreed with the statements about the current sample or the original population of interest: “Jamie is 93% confident that the mean monthly rent for the 100 selected students at HTSU is between $705 and $793”, and “Jamie is 93% confident that the monthly rent for all students at HTSU is between $705 and $793.” The participants were all able to identify that these statements were not about the value of the actual parameter of interest, the mean monthly rent for all students at HTSU. The statement with mixed agreement among the participants was a statement derived from a documented (mis)conception about the belief that the confidence interval represented the range of the sample statistics: “Jamie is 93% confident that the mean monthly rent for repeated samples of 100 students at HTSU is between $705 and $793.” Aiden agreed with this statement, while Tiana, Diana, and Gabe were unsure about the statement. Diana was initially confused but ultimately decided that it was not correct. Gabe, on the other hand, stated that 93% of samples should be within and 93% of confidence intervals should overlap the [actualized] interval. Tiana originally stated that the statement was incorrect but reversed that statement after answering Task 2, Jamie’s Colleague.

Table 8

Statements from Interview 2, Task 1, Question 1: Interpretations of Confidence Intervals

	Intro	Intermediate			Capstone				Graduate			Total
	Tiana	Kiara	Aiden	Logan	Diana	Emma	Brody	Gabe	Joel	Liam	Jace	Agreed
Jamie is 93% confident that the actual mean monthly rent for all students at HTSU is between \$705 and \$793. [a]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	10 of 11
Jamie is 93% confident that the actual mean monthly rent for all students at HTSU falls within \$705 and \$793.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	10 of 11
Jamie is 93% confident that the mean monthly rent for the 100 selected students at HTSU is between \$705 and \$793.	No	No	No	No	No	No	No	No	No	No	No	0 of 11
Jamie is 93% confident that the monthly rent for all students at HTSU is between \$705 and \$793.	No	No	No	No	No	No	No	No	No	No	No	0 of 11
Jamie is 93% confident that the mean monthly rent for repeated samples of 100 students at HTSU is between \$705 and \$793.	Maybe	No	Yes	No	Maybe	No	No	Maybe	No	No	No	1 of 11 (Yes) 3 of 11 (Maybe)

[a] notates the correct statement

Interpretation of Confidence Level

Interview 1, Question 4 and Interview 2, Task 1, Question 2 provided talking points for participants to express their current knowledge and understanding of the interpretation of confidence level. Interview 1, Question 4 stated: “Jamie and Alex cannot remember what the 95% represents in the calculation and the interpretation. How would you remind them what the 95% represents?” Interview 2, Task 1, Question 2 provided participants with five statements about possible interpretations for the confidence level. The participants’ answers to these questions could be categorized by general themes: long-run interpretation, coverage probability, and percentage. As they were answering the questions about the interpretation of a confidence level, participants could make statements that fell into multiple themes. As such, the categories of themes were not mutually exclusive. Rather, they contained any participant who made reference to that theme within their answers. Categorization of participants by themes are summarized in Table 7.

Long-Run Interpretation

The first category contained responses from Jace, Diana, Liam, Emma, Brody and Tiana who all implied a long-run interpretation of the confidence level. Participants in this category attempted (successfully and unsuccessfully) to interpret the confidence interval using the long-run interpretation: confidence level% of all possible samples of size n from a population will produce confidence level% confidence intervals that capture the actual value of the parameter. Diana, Liam, Emma, and Brody’s responses were similar to statements traditionally considered correct long-run interpretations. Tiana and Jace both attempted to discuss the long-run interpretation of the confidence level, but did not produce a response that would be considered a correct interpretation. Brody's statement demonstrated the types of correct long-run interpretation responses within this category (namely: Diana, Liam, Joel, and Emma):

> Yeah, so I'd say 95% represents, obviously the confidence level, but the fact that if you were to construct n 95% confidence intervals with the same population, approximately 95% of those intervals would contain the true parameter, p, p_j [Brody had previously defined p_j to represent the actual proportion of songs added by Jamie]

Jace discussed the long-run interpretation as replicates of many experiments: "if we doing replicates, many experiments, there are 95% chance the true parameters would fall into these confidence interval." Jace's statement differed from the others because he talked about the probability associated with a random interval. It was not necessarily clear from this or other statements made during his interview, whether Jace was referring to a random interval or an actualized interval. The distinction between a random or an actualized interval is important because Jace's interpretation is only correct if he was referring to a random interval. As Tiana attempted to think through the correct interpretation[6] of the confidence level, she spoke of recalling a similar statement from her introductory course: "I don't think we do that [indicating the second 93%]. I think a 93 just stays here. So, you would just have confidence interval." This statement suggested she had memorized and could identify the correct interpretation without connecting it to the underlying curricular concepts. The other participants in this theme category were

[6] For reference: "Approximately 93% of all samples of size 100 from the HTSU student body will produce 93% confidence intervals that capture the actual mean monthly rent for all students at HTSU."

able to discuss the long-run interpretation with more clarity than Tiana. Her inclusion of this interpretation in her response places her response in this theme but with the distinction that it was not as clear of a conceptualization as those held by the other participants in this theme category.

Coverage Probability

In addition to the long-run interpretation of the confidence level, many of the participants continued their discussion of the confidence level by including information related to the coverage probability. Participants in this category could discuss the following ideas related to coverage probability: 1) proportion of statistics within the sampling distribution related to the confidence level and 2) the relationship between the width of the confidence interval and the confidence level. This category contained the responses from the following participants: Liam, Brody, Logan, Jace, Joel, Emma, and Kiara. These participants can be grouped into three non-mutually exclusive sub-themes: 1) deeply connected coverage probability, 2) relation to width of the interval, and 3) compromise between the width of the confidence interval and the confidence level. These sub-themes are described below.

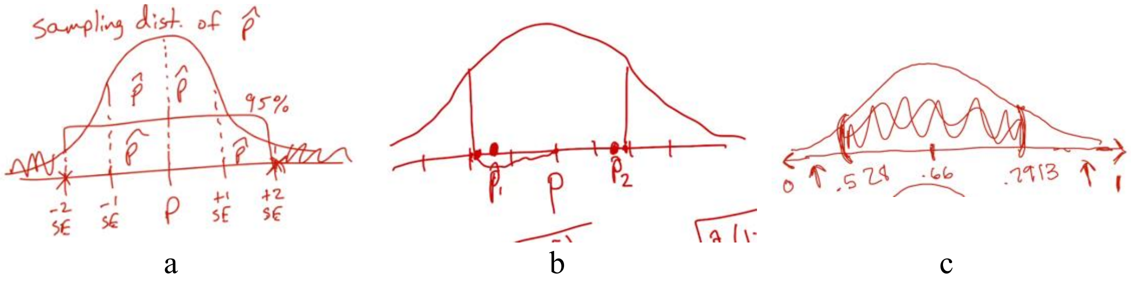
The first sub-theme is deeply connected coverage probability, which contained responses by Liam and Brody. They discussed the confidence level as relating to the proportion of sample statistics within the theoretical sampling distribution that will produce samples that would capture the actual value of the parameter (Figure 7a). Brody explains this idea as:

> So, when we find a $\hat{p}$, right, … So each of these, so let's say we choose a $\hat{p}$ [places $\hat{p}$ on the normal distribution in Figure 7b]. Right? So, there's our $\hat{p}$, we

> know that that $\hat{p}$ is going to have … a standard error, right that based on that standard error that comes from that $\hat{p}$. Right? We're going to create a confidence interval. … Then 95% of your $\hat{p}$ are going to be such that the spread of that interval [draws the line indicating the distance between the left vertical line and p on Figure 7b], right, contains p and so there is 95% of all possible $\hat{p}$ contain p on their interval. And so, I guess that's the best way I can say to explain that.

Figure 7

Drawing of the Confidence Level on the Sampling Distribution for Sample Proportions



a b c

The second sub-theme, *relation to width of the interval*, relates the width of the confidence interval to the confidence level. Logan, Jace, Joel, Emma, and Kiara discussed the idea of coverage probability by explaining that the confidence level is related to the width of the interval. Emma and Joel explained the connection between the confidence level and the confidence coefficient and its effect on the margin of error. Emma described this as:

> I would say 95 represents how much margin of error we're putting around the sample proportion, we know like a 99% confidence interval would be wider than a 90% confidence interval. So, the 95 is directly related to the z*, the critical value that I used. And it [referring to the 95%] represents the chance that our true proportion is within the interval.

Kiara, on the other hand, talked about the relationship between the confidence level and the width of the confidence interval, but with less concern about the compromise between confidence and width indicating a lesser understanding of coverage probability and confidence level. Specifically, she stated:

> So, 95 percent, or 100 minus a [meaning α], but I cannot remember what a [meaning α] is. And I don't think I'd be able to provide them with a 95% represents specifically, my big takeaway between the different percentages is that a 95% confidence interval is going to be smaller than a 95% [corrects herself by saying 90%] confidence interval.

The final sub-theme, *compromise between the width of the confidence interval and the confidence level*, contained responses from Logan, Joel and Jace. They connected the idea of coverage probability to the natural compromise that statisticians make when determining an appropriate confidence level: strength of confidence and precision of the confidence interval through the width of the confidence interval. Logan described this as:

> But it is kind of a sweet spot and having a narrow enough interval where, you know, it's um people can make interpretations off of it. So, if we were to use a 90% interval would be smaller, but it might not contain the true proportion we're looking for. And then if we went, we had a higher confidence level the interval, it would just be too big. And so I think you were kind of chosen 95% because it's that sweet spot.

These statements were mostly correct. Brody and Liam appeared to have a deeper understanding of the coverage probability than Jace, Joel, Emma, Kiara, and Logan. When Emma discussed the ideas of coverage probability, she consistently drew a picture similar to the one in Figure 7c. It is also not clear from the statements made by Jace, Joel and Logan whether they have the depth of connection to the sampling distribution that Brody and Liam demonstrated.

Percentage

The next category contained responses in which the participants interpreted the confidence level as a percentage. The percentage could refer to the long run-frequency ("the percentage of time" as the participants referred to this concept) the actual value of the parameter is between the upper and lower bound of the interval (both actualized and random) or the long-run frequency (percentage of time) the data points are within the interval. Kiara, Emma, Aiden, Logan, Jace, Joel, Gabe and Tiana's responses fell in the *Percentage* theme. Their responses, however, can be grouped into the following sub-themes: 1) percentage of time a value is within the interval, 2) percentage of time a value is within the actualized confidence interval.

Kiara, Emma, Aiden and Tiana's responses were grouped into the first sub-theme, percentage of time the actual value of the parameter is within the confidence interval. Emma and Aiden both implied that 95% of the time the actual value of the parameter is within the range, as demonstrated by Emma: "95% of the time, the true proportion songs that Jamie load onto the playlist would be somewhere within our interval." Tiana, expressed a similar idea but discussed the idea of the confidence level as the proportion of the actual value of the parameter that is within the confidence interval: "the range of where 95% of the true proportion of Alex's (sic) songs in the playlist." Finally, Kiara's response discussed the percentage of time the population data values are within the interval: "95% of the time, all the data points you pull from the population are going to fall within 95 of the 100 samples.” Emma and Aiden appeared to have a conception that combined the long-run interpretation with the idea that the level was the long-run frequency (percentage of time) a value (data points or actual value of the parameter) is within an interval. Kiara and Tiana demonstrated confounding ideas, perhaps indicating they had memorized the statements and lack conceptual grounding.

Over the course of the interviews, Logan, Emma, Jace, Joel, and Gabe discussed the interpretation of the confidence level as the *percentage of time a value was within the actualized interval*, the second sub-theme. Logan, Emma, Jace and Joel discussed the interpretation of the confidence level as long-run frequency (the percentage of time) the actual value of the parameter is within the confidence interval. Emma’s statement was an exemplar of this conception: “95% of the time, the true proportion songs that Jamie load onto the playlist would be somewhere within our interval.” Gabe described the interpretation of the 95% as the proportion of samples that will have results within the provided interval: “we're basically saying that we think that 95% of the samples from the population are going to show results that we saw before or going to have proportions that are within those confidence bands that we saw before.” The placement of percentage or probability within the actualized interval is problematic because the probability rests with the random process used to generate the formula for the actualized confidence interval. It is hypothesized that the responses in this theme generate the (mis)conceptions proportion of the population, probability of the actual value of the parameter being in the interval, and proportion of the sample statistics (see: Table 3).

Statements from Interview 2

As with the interpretations of the confidence interval, the diverse nature of the statements originally made by the participants when they were asked to discuss the interpretation of the confidence level might lead one to believe the participants would not all choose the same statements on Interview 2 as correct. As seen in Table 9, this is not the case. Liam did not receive these statements over the course of the three interviews he participated in and is not included further discussions in this section. All of the participants disagreed with the statements:

1. "Approximately 93% of the actual mean monthly rent for all students at HTSU is between $705 and $793."
2. "Approximately 93% of the mean monthly rent for the 100 selected students at HTSU is between $705 and $793."
3. "Approximately 93% of that the mean monthly rents for repeated samples of 100 students at HTSU is between $705 and $793."

All but Aiden disagreed with the statement: "Approximately 93% of the monthly rent for all students at HTSU is between $705 and $793." Aiden originally started to say that this type of statement would be wrong. As he continued his explanation, however, he stated that if the interval represented means, then the interval should also contain the individual data points before admitting that he may not be right. Aiden's response implied that he may be visualizing the range of sample means to be similar to the range of the original data. The statement in the five-statement slide gave the most diverse answers, which was the correct statement, was: "Approximately 93% of all samples of size 100 from the HTSU student body will produce 93% confidence intervals that capture the actual mean monthly rent for all students at HTSU." Three participants were not in favor of the double statement of the confidence level. Jace and Joel both disagreed with "approximately 93% of all samples…" and wanted to replace the 93% to 100% without providing an explanation of why all of the samples would produce intervals that capture the actual value of the parameter. Tiana wanted to remove the additional 93% without a replacement. Kiara flatly disagreed with this statement and eliminated the entire first part of the sentence, instead focusing on the interpretation of the confidence interval:

> 93% confident that the interval captures the actual mean monthly rent for all students. That to me is the proper definition. And then adding on approximately 93% of samples of size 100 is just like, like weird additional information, but I don't know how you could like come up with that, at least off the top of my head. I don't know.

Logan stated that the statement was wordy, and other participants disliked the second 93% but eventually agreed with the statement. This statement is more specific than the statement traditionally discussed as the interpretation: "Approximately 95% of confidence intervals will capture the actual value of the parameter." Many of the participants commented on the over-specification of this statement. Logan, Diana, Emma, Brody, and Gabe all agreed with this statement.

Table 9

Statements from Interview 2, Task 1, Question 2: Interpretation of Confidence Levels

	Intro	Intermediate			Capstone				Graduate			Total
	Tiana	Kiara	Aiden	Logan	Diana	Emma	Brody	Gabe	Joel	Liam	Jace	Agreed
Approximately 93% of the actual mean monthly rent for all students at HTSU is between $705 and $793.	No	No	No	No	No	No	No	No	No	N/A	No	0 of 10
Approximately 93% of the mean monthly rent for the 100 selected students at HTSU is between $705 and $793.	No	No	No	No	No	No	No	No	No	N/A	No	0 of 10
Approximately 93% of the monthly rent for all students at HTSU is between $705 and $793.	No	No	Maybe	No	No	No	No	No	No	N/A	No	1 of 10 (Maybe)
Approximately 93% of that the mean monthly rents for repeated samples of 100 students at HTSU is between $705 and $793.	No	No	No	No	No	No	No	No	No	N/A	No	0 of 10
Approximately 93% of all samples of size 100 from the HTSU student body will produce 93% confidence intervals that capture the actual mean monthly rent for all students at HTSU. [a]	Maybe	No	No	Yes	Yes	Yes	Yes	Yes	Maybe	N/A	Maybe	5 of 10 (Yes) 2 of 10 (Maybe)

[a] symbolizes correct response

Classification of Conceptions of Interpretations

The participants had diverse responses to the questions and analysis of statements about the interpretations of the confidence interval and the confidence level. Table 7 contains the summary of the categorizations of participants. In order to create classifications of conceptions of the interpretation of the confidence interval and the interpretation of the confidence level, I looked for patterns among the categories of conceptions across the participants of this study. Appendix M: Table 19, Table 20 and Table 21 contain the final groupings of the participants that based on the classifications that are discussed in this section.

Interpretation of the Confidence Interval

There were 4 four classifications of the interpretation of the confidence interval (see: Table 10): 1) Correct, 2) Not Probability, 3) Replicate of Experiment, and 4) Incompatible. There are four possible conceptualizations for responses about the interpretation of the confidence interval: 1) Correct, 2) Capture/Not Capture, 3) Long-Run and Probability, and 4) Confounding. Of these categories, two are potentially correct interpretations of a confidence interval: Correct and Capture/Not Capture. Having a correct understanding of confidence intervals does not necessarily require that an individual discusses both of these categories but should not discuss any of the other categories. According to some experts, the long run interpretation of the confidence level may be a correct interpretation of the confidence interval. Therefore, the *Correct* classification consists of individuals who conceptualize the interpretation of the confidence interval as at least one of the following categories with no other conceptions: 1) Correct, 2) Long-Run interpretation of the confidence level correctly, and 3)

Capture/Not capture. The second classification, *Not Probability*, contains individuals who use a correct interpretation of the confidence interval while focusing on the lack of probability. Individuals in this classification discuss the interpretation clearly in terms of not containing probability because of the capture/not capture scenario. The third classification is *Replicate of Experiment*, which includes individuals whose conceptualizations of the interpretation of the confidence interval consists of long-run interpretations through repeated sampling. The final category consists of individuals who present *Incompatible* statements by expressing confounding or contradictory statements when interpreting the confidence interval.

Table 10

Classification of Conceptualization Themes for the Interpretation of Confidence Intervals

		Correct	Not Probability	Replicate of Experiment	Incompatible
Confidence Interval	Correct	Yes			
	Capture/ Not Capture	Yes	Yes		
	Long-run/ Probability		Yes	Yes	
	Confounding				Yes

Interpretation of the Confidence Level

There were few similarities within the categorizations of the interpretations of confidence level despite only having three possible categories for confidence level responses: long-run interpretation, coverage probability, and percentage of time/sureness/interval. There are 4 classifications of conceptualization patterns (see: Table 11): 1) Correct, 2) Probability and Width, 3) Parallel, and 4) Percentage. For the interpretation of a confidence level, two of the three categories could represent participants with correct conceptions of the interpretation: long-run and coverage probability. Like the interpretation of a confidence interval, mentioning both categories was not necessary for a fully correct understanding of the interpretation of a confidence level. Individuals in the *Correct* classification are able to are demonstrate and discuss robust knowledge of both the long-run interpretation of the confidence level and discuss the coverage probability both in terms of confidence coefficients and sampling distributions. The next classification is *Probability and Width* which contains individuals who conceptualize the interpretation of the confidence level with components of the coverage probability conceptualization while interpreting the confidence level as percentage of time or sureness of the actual value of the parameter is within the upper and lower bounds of the confidence interval. The third classification, *Parallel*, consists of individuals who conceptualize the interpretation of the confidence level with all the three conceptualization categories: 1) using the long-run interpretation of the confidence level, 2) discussing components of the coverage probability, and 3) interpreting the confidence level as the probability of the actual value of the parameter is within the upper and lower bounds of the confidence interval. The final classification is *Percentage*, which includes individuals who conceptualize the interpretation of the confidence level strictly as a percentage of time.

Table 11

Classification of Conceptualization Themes for the Interpretation of Confidence Level

		Correct	Probability and Width	Parallel	Percentage
Confidence	Long-run	Yes		Yes	
Level	Coverage Probability	Yes	Yes	Yes	
	Percentage		Yes	Yes	Yes

Similarities and Differences in Concept Images (R2)

This section discusses different potential dimensions that emerged as areas of similarities and differences in undergraduate and graduate students' concept images of the interpretations of confidence intervals and the interpretation of confidence levels. Initial analysis indicated there existed five possible dimensions of an individual's concept image that may affect a person's interpretations: 1) the meaning of the word confident, 2) the understanding of the confidence level, 3) understanding of probability, 4) random process (Jamie's Colleague), and 5) visualization of the confidence interval. This dissertation study focuses on the first two dimensions: the meaning of the word confident and the understanding of the confidence level, leaving the other three to future directions. Similar to the analysis of general interpretations provided in the previous sections, the participants were categorized based on themes stated during the first two interviews. As with the previous sections, these categorizations are not mutually exclusive as participants often stated contradictory or parallel conceptualizations.

Concept Image Dimension: Confident

During the analysis of the participants' conceptualization of the interpretation of the confidence interval, it became apparent that participants held differing meanings for the word confident. Thus, the definition of the word confident may be a potential dimension of each participant's concept image where the personal concept definition differs from the formal concept definition. The participants' personal definitions for the word confident can be categorized into three themes: 1) quantifiable measure, 2) sureness/belief, and 3) confidence/chance/probability.

Quantifiable Measure

The first theme, *quantifiable measure*, is invoked when a participant referred to the use of the word confident as a way to quantify the estimate or the statistical process being used, revealing a somewhat vague definition for the word confident. There are two sub-themes to the responses in this category: 1) estimating the actual value of the parameter well and 2) identifying the statistical process being used. Brody, Liam, Diana, and Gabe defined the word confident as a quantifiable measure of the estimate (the confidence interval) was doing what it was supposed to be doing (*estimating the actual value of the parameter*). It is not immediately clear what these four participants meant by the "quantifiable measure" (Brody), "doing what it's supposed to" (Liam) or "representing whatever you're trying to estimate" (Diana). Brody described this as: "statistically speaking that like it's a quantifiable measure of how certain you are that your estimate is correct, essentially. … So, yeah, I would say that it's a quantifiable measure of how good you feel about your estimate." Brody explained further how the confidence interval estimated the actual value of the parameter and was not flustered by the request to define the word confident. Liam, on the other hand, was not able to explain further what "doing what it's supposed to" meant, beyond estimating the actual value of the parameter. Diana, instead, discussed the measure of strength as "how strong your value compares to the true value." It is evident, however, that Liam, Brody and Diana had not thought about how to define the word confident, as used in the interpretation. All three were clear, however, that the definition of confident does not include the word probability.

The second sub-theme contained participants whose responses indicated the choice of the *word confident referred to the statistical method being used*. Brody, Liam, Gabe, and Logan preferred the word confident because of its recognition of the statistical procedure (confidence interval) being used. Brody, Liam, Logan, and Gabe described the choice of the word confident in the interpretation of the confidence interval because it represented the statistical procedure that was being used. Logan stated: “confident implies that we are estimating, and we are in the percentage level of confidence in that estimate, which is why I like the word.” It is unclear if either of these meanings are helpful definitions for confident, but most participants in this category (4 of 5) correctly interpret the confidence interval and three participants correctly interpreted the confidence level.

Sureness/Belief

The next theme categorized the responses that defined the word confident as being equivalent to the words sure and/or belief or described the word confident as a measure of sureness and/or belief. Unlike the previous categories of response themes, all of the participants in this group provided similar responses. Liam, Aiden, Kiara, Gabe, and Tiana defined confident as equivalent to sureness and/or belief. Liam and Gabe implied belief in their definition: “Jamie has, like 93% belief that the actual mean falls within those two numbers (Liam).” Aiden referred to confident as “an arbitrary declaration,” implying belief or likelihood. Kiara stated that it implied “more likely than not… you’re almost super certain… way more likely than not, like a 51%.” For Aiden, Kiara, and Tiana, the equivalence appeared to be a slippery slope towards defining the word confident in terms of probability. These three stated explicitly that confident does not imply probability, but their discussions in the two interviews suggested they have come to understand there is probability associated with the [actualized] interval.

Confidence/Chance/Probability

The next theme discusses different conceptions of the *equivalence among the words: confidence, chance, and probability*. Probability and chance are assumed to be equivalent words in this dissertation study. Within the context of interpreting confidence intervals, confidence is not an equivalent word choice to probability or chance. There are two distinct sub-categories that pose potentially different conceptualization issues: 1) basic confusion about the equivalence of the three words and 2) assumed equivalence of the three words. Those confused about the nonequivalence of the words may be developing an understanding of the underlying concepts behind the difference in the word choice of confident versus chance and probability. Those who believed the three words were equivalent may be demonstrating a different concept image. They will be discussed together in this section but grouped separately for further analysis.

The first sub-group contains those participants who demonstrated *confusion over the lack of equivalence of the words: probability, chance and confidence.* Aiden, Kiara, Tiana, Liam, and Joel made statements classified as confusion about the word choice of confidence over probability and chance. There were two types of confusion with respect to the equivalence of the three words: 1) initial confusion that has been 'corrected,' and 2) contradictory statements about the equivalence of the words across the interviews. Joel and Liam struggled previously with understanding the choice of confident. Liam's struggle started in AP statistics, but he was able to convince himself using the capture/not capture scenario that confidence and probability were not equivalent. Joel, on the other hand, simply learned the word confident (in both English and Chinese): "to be honest, I don't know [referring to the difference between probability and confidence]. I have the same question before, but I did not find the answer. Yeah. And still, in my heart, yeah. Since everyone talks, the official name, confidence interval. Yeah, I just follow that." Aiden, Kiara, and Tiana originally interpreted the definition of confident as equivalent to chance, but eventually stated the two words were not equivalent. Aiden immediately contradicted himself during the first interview by stating that he felt chance was never the right word to use. Kiara and Tiana, however, stated they were not equivalent during the second interview.

The second sub-category contained those participants who stated that confident was *equivalent to chance and probability*. During both interviews, Logan, Emma and Jace stated explicitly that the words confidence, sure, chance, and probability were equivalent. In Jace's case, there may have been a translation issue. The subtle difference between these words in the English language is difficult to understand for native speakers. The difference in meaning, statistically, is quite large but may be difficult for non-native speakers to understand.

The participants, in general, agreed on correctness of the statements on Interview 2, Task 1, Question 2 that had the same statement but with sure, probability, and chance substituted for confident (see Table 12). Aiden, Diana, and Liam participated in Interview 2 prior to the statement with the word chance being added to the protocol. Logan, Emma, and Jace agreed with the statements, implying an equivalence among the words confident, chance, and probability. The other 8 participants disagreed with the statements that contained probability and chance. Brody and Joel were the only participants who disagreed that sure was equivalent to the word confident. The rest of the participants agreed with the statement.

Table 12

Statements from Interview 2, Task 1, Question 3: Confident

	Intro	Intermediate			Capstone				Graduate			Total
	Tiana	Kiara	Aiden	Logan	Diana	Emma	Brody	Gabe	Joel	Liam	Jace	Agreed
Jamie is 93% *confident* that the actual mean monthly rent for all students at HTSU is between $705 and $793.[a]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	11 of 11
Jamie is 93% *sure* that the actual mean monthly rent for all students at HTSU is between $705 and $793.	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	9 of 11
Jamie has a 93% *probability* that the actual mean monthly rent for all students at HTSU is between $705 and $793.	No	No	No	Yes	No	Yes	No	No	No	No	Yes	3 of 11
Jamie has a 93% *chance* that the actual mean monthly rent for all students at HTSU is between $705 and $793.	No	No	N/A	Yes	N/A	Yes	No	No	No	N/A	Yes	3 of 8

[a] notates the correct answer

Concept Image Dimension: Confidence Level

The second dimension of potential similarities and differences among the concept images of the participants was the concept of the confidence level. Within this concept, a depth of knowledge appears to be the biggest difference among the participants. For instance, some participants viewed the confidence level simply as a quantifier: a value that was used to find the confidence coefficient needed to calculate the interval. On the other end of the spectrum, some participants were able to explain the confidence level as the coverage probability, including its connection to the long-run interpretation of the confidence level. The categorizations of the conceptions of the confidence level, like the previous categorizations, are reliant on what was discussed during the interview. Some participants may have conceptions of the confidence level that fit into other categories but were not demonstrated during the interviews. The responses were categorized into four themes and are explained in the following sub-sections: 1) Confidence Coefficient, 2) Relation to Width, 3) Probability and Accuracy, and 4) Coverage Probability.

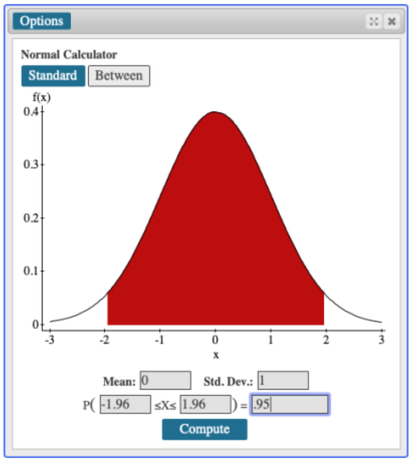
Confidence Coefficient

All of the participants, at some point, were able to identify the confidence coefficients for the particular confidence level of interest. Three participants, however, appeared to have only a *procedural understanding of the confidence level*. Tiana, Kiara, and Diana all appeared to view the confidence level as a way to find the values they needed to calculate the confidence interval. Tiana's explanation is particularly telling in that she recalled the exact procedure for finding the confidence coefficient using the normal calculator in StatCrunch (Figure 8):

> like there was some plugging into StatCrunch, like you probably could do it by hand, somehow. You need to figure out what your mean is, and I remember the standard deviation thing. And then you put like, .95 somewhere in the thing…. I think you look at something like you have an X here. And there were boxes. And that tells you the answer.

Figure 8

StatCrunch Normal Calculator Applet



Mean: 0 Std. Dev.: 1

P(-1.96 ≤X≤ 1.96) = .95

Note. From *Normal Calculator Applet*, by StatCrunch, 2019, (www.statcrunch.com), 2019 by Pearson. Reprinted with permission.

Similarly, Kiara did not explain the confidence level beyond its use to calculate the confidence interval. Diana's explanation was a bit different because she had had more advanced statistical courses than either Tiana or Kiara. To Diana, the strength in the confidence level was based in the distribution, stating that the confidence comes from the calculations within the distribution:

> I think that's captured in whatever the, the critical value is. Which is kind of why we use the critical values. Because I think they've kind of been established to give

> that level of confidence as long as you can prove that. Like you've met the minimum criteria for whatever distribution you're using.

The difficulty is that Diana admitted to not remembering the connection between the sampling distribution and the population. Therefore, it appears her conceptualization of this distribution was not connected to the coverage probability, essentially limiting her understanding in a way that is similar to Tiana and Kiara.

Relation to Width

This category is similar to the confidence coefficient category, but it contained response from participants who specifically *related the confidence level to the width of the interval*. The participants in this group were Kiara, Aiden, Logan, Jace, Joel, Emma, and Aiden. There were two related sub-themes (similar to the sub-themes in the Coverage Probability categorization of the interpretations of the confidence interval): 1) the relationship between the confidence level and the margin of error and 2) concern about the precision of the interval. Kiara, Joel, Emma, and Aiden talked about the immediate *relationship between the confidence coefficient and the margin of error*. Emma described this as: "I would say 95 represents how much margin of error we're putting around the sample proportion, we know like a 99% confidence interval would be wider than a 90% confidence interval." Logan, Joel and Jace expressed concern about picking the best confidence level that produces a "sweet spot," as Logan stated, between the width of the interval and measure of confidence (*concern about the precision of the interval*).

Probability and Accuracy

The next theme contained responses conceptualizing the confidence level in probabilistic terms. While the confidence level is equivalent to the coverage probability, the coverage probability is the probability the confidence interval estimator captures the unknown value of the parameter. Using the coverage probability as the probability the actualized interval captures the actual value of the parameter could lead to students defining the word confident as equivalent to probability or interpreting the interval in a probabilistic way. Emma, Gabe, Jace, Aiden, Tiana and Joel provided responses that were categorized into this theme. There were two sub-themes of this category: 1) probability the actualized interval captured the actual value of the parameter and 2) accuracy that the interval captured the actual value of the parameter. Emma, Gabe, Aiden and Jace all expressed that the confidence level referenced the *probability the actualized interval captured the actual value of the parameter*. Tiana and Joel both described the confidence level as the *accuracy of the interval*. Tiana defined the word accurate as: "getting the true proportion or the true mean of the population and then representing what it's supposed to." Joel described accuracy using the long-run interpretation of the confidence level and significance.

Coverage Probability

The final category for responses for the conceptualization of the confidence level is *Coverage Probability*. Brody, Liam, Gabe and Emma explained their conceptualization of the confidence level in ways were categorized in this theme by describing the confidence level in a more deeply connected way and relating the confidence level to the sampling distribution. Some participants continued this explanation by explaining the meaning of the long-run interpretation of the confidence level. There were four major components demonstrated by participants in this category: 1) the confidence level is related to the proportion of statistics within the middle region of the sampling distribution centered at a value of the unknown parameter (see Figure 7a), 2) explaining the choice of the margin of error as half the width of the confidence level% region of the sampling distribution (Figure 7b), 3) connecting the proportion of statistics within the middle region of the sampling distribution to the long-run interpretation of the confidence level, and 4) relating the long-run interpretation with the probability of randomly selecting a sample from the population. To be part of this category, participants must have discussed a combination of the four components. Brody explained the confidence level using all four components of the coverage probability. As explained in Brody's summary (see: Appendix L, Capstone Students, Brody), the first three components were a normal part of Brody's conceptualization. Brody made the connection to the fourth component, during the interview. Liam explained his understanding of the confidence level using the first three components. He did not have the connection between the interval estimator (random interval) and the probability of randomly selecting a sample from the population. Liam disregarded any connection between the underlying probability of a random interval and an [actualized] interval. The final two participants in this category discussed the region of the sampling distribution that contained the confidence level proportion of statistics. Gabe and Emma described this region as having endpoints that were related to the currently calculated confidence interval rather than abstractly as Liam and Brody did. Gabe, however, described the confidence interval by explaining the choice of the margin of error (component 2) as the fixed with of the actualized interval and the long-run interpretation of the confidence level (component 3). Table 13 summarizes the components used by each participant. All of the four participants whose responses were categorized as Coverage Probability interpreted the confidence interval correctly. Brody and Liam were the only ones to use only correct conceptualizations of the interpretation of the confidence level. Appendix M: Table 19 lists the participants by statistical experience and how their conceptualizations of confidence and confidence level were categorized.

Table 13

Summary of Components Used When Discussing Coverage Probability

Participants / Components	Brody	Liam	Gabe	Emma
Proportion of statistics	Yes	Yes	Yes	Yes
Margin of Error	Yes	Yes	Yes	
Long-Run	Yes	Yes	Yes	
Probability	Yes			

Classification of Confident and Confident Level

Patterns across the participants' conceptualizations of confident and the concept of confidence level produced the classifications presented in this section. Each dimension is addressed separately. Appendix M: Table 22 and Table 23 present the application of these classifications to the participants in this dissertation study. This section, instead, focuses on describing the classifications as a model for categorizing individuals' conceptualizations of the word confident and the concept of confidence level.

Concept Image Dimension: Confident

Based on the responses, there were originally four non-mutually exclusive categories: 1) Quantifiable Measure, 2) Sureness/Belief, 3) Confusion of Confidence, Chance, Probability and 4) Equivalence of Confidence, Chance, Probability. These categories did not necessarily represent the correct statistical definition of the word confident. The category closest to the statistical definition is confident as a "quantifiable measure." This alone, however, is not sufficient to say an individual has a correct definition of the word confident. I argue that this symbolizes a lack of definition because of its missing connection to the random process. Sureness/Belief may be correct meanings for the word confident, with the caveat that the individual does not imply that sure and belief are equivalent to probability. Finally, confusion and/or assumed equivalence of the words confidence, chance, and probability are most likely evidence of some non-productive conceptions. Based on conceptualization patterns, four classifications were evident (see: Table 14): 1) Quantifiable Measure, 2) Belief, 3) Towards Probability and 4) Probability. Individuals who can be classified as having *Quantifiable Measure* as their conceptualization of confident only discuss the word confident as a quantifiable measure/ doing what it's supposed to. Individuals who are classified with the *Belief* conceptualization define the word confident as it being a quantifiable measure, equivalent to sureness and belief, and/or expressed confusion about the words confidence, chance, and probability. *Towards Probability* is a classification that represents those who appear to view the word confident as a belief but are expressing confusion about the equivalence of the words confident, chance, and probability. The final category is *Probability*, which contains individuals who conceptualize the word confident as equivalent to the words chance and probability.

Table 14

Classification of Conceptualization Themes for Confident

		Quantifiable Measure	Belief	Towards Probability	Probability
Confidence	Quantifiable Measure	Yes	Yes		
	Sureness/ Belief		Yes	Yes	
	Confusion: Conf/Chance			Yes	
	Equival: Conf/Chance				Yes

Concept Image Dimension: Confidence Level

The participants displayed varying conceptualizations of the concept of confidence level. There were four categories for participants' conceptions of the concept of confidence level: 1) Confidence Coefficient, 2) Relation to Width, 3) Probability/Accuracy, and 4) Coverage Probability. Of these categories, the Confidence Coefficient, Relation to Width, and Coverage Probability conceptualizations potentially hold productive conceptions of the confidence level. These conceptualizations deepen in connections from knowing the procedure of finding the confidence coefficient, to relating it to the width of the interval to the more connected conceptualization of coverage probability. This means that individuals with conceptualizations that align with Confidence Coefficient and Relation to Width may not have developed a robust correct conceptualization of the concept of confidence level. Participants' conceptualizations varied greatly, producing many different combinations of categorizations. There were three classification categories for the concept of confidence level (see: Table 15): 1) Coverage Probability, 2) Developing and 3) Surface Level. *Coverage Probability* classification contains only conceptualizations of the concept of confidence level as a combination of the components of Coverage Probability (see Table 13) that may be required to understand the concept of confidence level robustly. The *Surface Level* category is also distinct in that the conceptualizations in this category appear to be procedural in nature: the confidence level is only used to find the confidence coefficient or suggest belief in the interval. The *Developing* category indicates individuals who had varying conceptualizations of the concept of confidence level. This part of an individual's concept image appears to be developing because there exists more connections beyond a

Surface Level understanding but are missing strictly robust understandings that are required for the Coverage Probability category.

Table 15

Classifications of Conceptualization Themes for Confidence Level

		Coverage Probability	Developing	Surface Level
Confidence Level	Confidence Coefficient		Yes	Yes
	Relation to Width		Yes	
	Probability/ Accuracy		Yes	
	Coverage Probability	Yes	Yes	

Classification of the Concept of Confidence Intervals

The final result presented in this dissertation study is a classification system for conceptualizations of the entire concept of confidence intervals. Within the responses for the interpretation of the confidence interval, there were four main themes: 1) Correct Interpretation, 2) Capture/Not Capture, 3) Long-Run Interpretation and Probability, and 4) Confounding Ideas. The interpretations of the confidence level had three main themes: 1) the Long-Run Interpretation, 2) Coverage Probability, and 3) Percentage. The word confident contains four main themes: 1) Quantifiable Measure, 2) Sureness/Belief, and 3) Confusion of Confident, Probability, and Chance, 4) Equivalence of Confident, Probability, and Chance. The concept of confidence level contains four themes: 1) Confidence Coefficient, 2) Relation to Width, 3) Probability and Accuracy, and 4) Coverage Probability. Participants discussed a theme in different ways, which resulted in several categorizations having sub-themes (see Table 16). These categories of conceptualization themes for each participant grouped by statistical course can be found in Appendix N: Table 24.

Table 16

Table of Themes for the Interpretations, Confident, and Confidence Level

	Category	Sub-Theme
Interpret Confidence Interval	Correct	Correct Hypothesis Formulaic
	Capture/Not Capture Long-run Interpretation and Probability Confounding Ideas	
Interpret Confidence Level	Long-run	Correct Repeated Procedural
	Coverage Probability	Connected Probability & Width Compromise
	Percentage	Value Actualized
Confident	Quantifiable Measure	Estimating Process
	Sureness/ Belief Confusion of Confident/Chance Equivalent of Confident/Chance	
Confidence Level	Confidence Coefficient Relation to Width Probability and Accuracy Coverage Probability	

Classification Categories for the Concept of Confidence Intervals

Patterns of categorization of conceptualization themes across the four areas analyzed in this dissertation study were used to create classifications for the entire concept of confidence intervals. There were four participants who clustered into two distinct pairs: 1) Kiara and Tiana, 2) Brody and Liam. Kiara and Tiana demonstrated the most surface level knowledge and made the most inconsistent statements across the two interviews. On the other extreme, Brody and Liam held conceptualizations that were the most connected and conceptually developed. These two pairs provided examples of the range of knowledge and connections that appeared in this dissertation study and could be considered the anchor points between surface level understanding of confidence intervals and deep understanding of confidence intervals. These pairs were classified as *Inconsistent* and *Connecting*, respectively. *Inconsistent* individuals represent those who: 1) demonstrate confounding interpretations of the confidence interval, 2) view the interpretation of the confidence level as a percentage of time, 3) define confidence in terms of sureness and expressed confusion about the difference in the words confidence and chance, and 4) demonstrate procedural understanding of the confidence level. *Connecting* individuals conceptualize: 1) the interpretation of the confidence interval correctly and with capture/not capture explanations, 2) the interpretation of the confidence interval using the long-run interpretation and coverage probability, 3) the word confident using quantifiable measure reasoning, and 4) the confidence level as highly connected within the sampling distribution. The remaining participants held very different patterns of conceptualizations for the four areas of analysis. Among the remaining seven participants, one pair of participants appeared to be similar to each other and different enough from the others to be considered a core pair: Logan and Joel. These two participants form the core conceptualizations of the final classification category. *Developing* included the following conceptualizations: 1) interpreting the confidence interval as a long-run interpretation and as a probability, 2) interpreting the confidence level as a percentage, and 3) relating the confidence level to the width of the confidence interval. The place of greatest difference between the pair was the definition of confident.

The final three groupings of conceptualization themes and classification categories can be found in Table 17. This table also includes shading to indicate whether the reasoning that was demonstrated within the conceptualization theme was correct, developing, or incorrect. Correct reasoning, red shaded categories, considered statements and ways of thinking that are considered correct. For instance, the sub-theme Interpret Confidence Interval – Correct – Correct, contained responses that used the traditional interpretation of the confidence interval. The categories shaded tan indicate developing conceptualizations. These conceptualizations were potentially incorrect statements but may indicate 1) a developing idea that could become an area of productive struggle or reasoning, or 2) a correct statement that may or may not be a conceptualization that helps individuals develop deep understanding of the concept of confidence intervals. For instance, a confounding statement for the interpretation of the confidence interval is a statement that combines two ideas into one statement, as Tiana's statement did: "We are 95% confident that, if we took 100 sample tests [indicating repeated samples], the true proportion would fall between this range." This may or may not be a path of development that will help produce deep understanding. Statements that were grouped as hypothesis test or formulaic within the Correct category of the Interpretation of the Confidence Interval were technically correct statements but may or may not be paths of productive reasoning for the interpretation of the confidence interval. The black shaded categories represent incorrect reasoning, such as the interpreting the confidence level as the probability the actual value of the parameter is within the upper and lower bounds of the confidence interval. This is an incorrect statement that does have the potential for developing non-productive paths to deeply understanding confidence intervals. Finally, there were categories that contained both developing and incorrect conceptualizations, shaded gray. For instance, conceptualizing the confidence level using a response categorized as probability and accuracy contained both developing (accuracy) and incorrect (probability) reasoning.

Table 17

Classification of Conceptualization Themes for the Concept of Confidence Interval

	Category	Sub-Theme	Inconsistent	Developing	Connecting
Interpret Confidence Interval	Correct	Correct			🟥
		Hypothesis			
		Formulaic			🟫
	Capture/Not Capture				🟥
	Long-run and Probability			▒	
	Confounding Ideas		🟫		
Interpret Confidence Level	Long-run	Correct			🟥
		Repeated			
		Procedural	🟫		
	Coverage Probability	Connected			🟥
		Prob & Width	🟫		
		Compromise		🟫	
	Percentage	Value	🟫		
		Actualized		⬛	
Confident	Quantifiable Measure	Estimating			🟫
		Process		🟫	🟫
	Sureness/ Belief		🟫		🟫
	Confusion of Confidence		🟫	🟫	🟫
	Equivalent of Confidence			⬛	
Confidence Level	Confidence Coefficient		🟫		
	Relation to Width		🟫	🟫	
	Probability and Accuracy		▒	▒	
	Coverage Probability				🟥

Note. Red indicates correct reasoning. Tan indicates developing reasoning. Black indicates incorrect reasoning. Gray indicates categories that had both developing and incorrect reasoning.

Application of the Classification Categories to the Participants

As stated above, there are three core pairs for each classification: 1) Kiara and Tiana for the *Inconsistent* group, 2) Logan and Joel for the *Developing* group, and 3) Liam and Brody for the *Connecting* group. The remaining seven non-core-pairs participants share similarities and differences with each other and were placed into the three classifications based on how similar they were to the core pair. Figure 9 portrays the final classification of all of the participants for their conceptualizations of the concept of confidence intervals. The figure shows the following groupings: 1) Inconsistent (blue shaded) - Tiana, Kiara, and Aiden, 2) Developing (tan shaded) - Logan, Joel, Jace, and Emma, and 3) Connecting (red shaded) - Gabe, Liam, Brody, and Diana. It is important to note that there may be cross-category similarities among these seven non-core-pairs participants. To illustrate this, the diagram shows the participants as they were classified. The distance shown on the diagram is not meant as a measure of how closely aligned the participants are within a group. Rather, the placement of participants is meant to illustrate similarities of participants' conceptualizations both within and across classification categories. For instance, Emma and Gabe were classified into different groups, but as Table 18 shows they had similar conceptualizations for the interpretation of the confidence interval but held very different conceptualizations across the other three areas of analysis. In Figure 9, Emma and Gabe are shaded according to their classification, but are on the same side of the diagram indicating their similarity to each other but also to their classification. The next few paragraphs explain the rationale for the classification of the seven non-core participants into each classification category.

Table 18

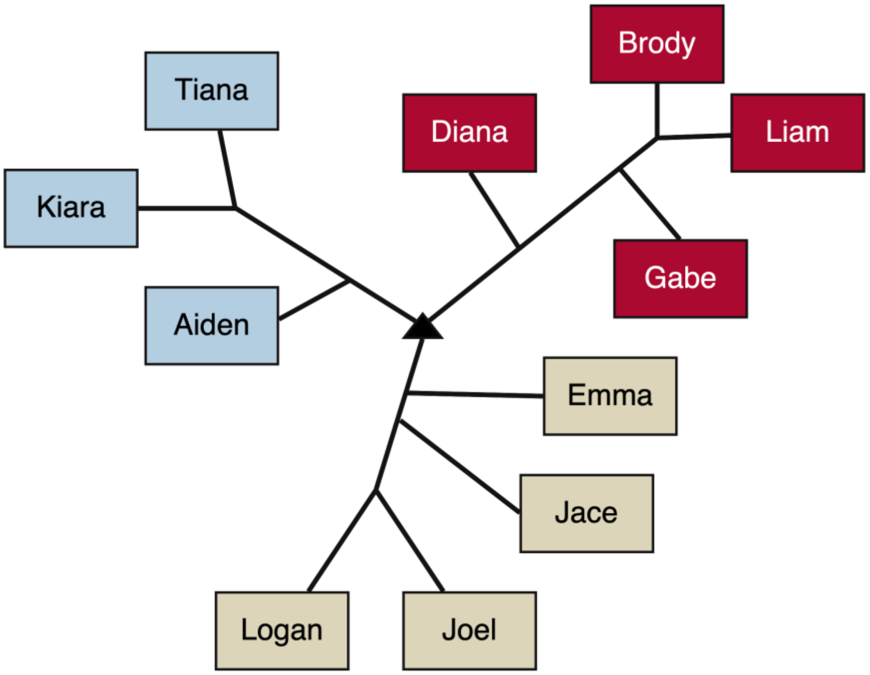
Classification of Participants' Conceptualization Themes for the Concept of Confidence Interval

			Inconsistent			Developing				Connecting			
	Category	Sub-Theme	Tiana	Kiara	Aiden	Logan	Joel	Jace	Emma	Gabe	Liam	Brody	Diana
Interpret Confidence Interval	Correct	Correct			🟥				🟥	🟥	🟥	🟥	
		Hypothesis							🟫	🟫			
		Formulaic									🟫	🟫	
	Capture/Not Capture							🟥	🟥	🟥	🟥	🟥	🟥
	Long-run and Probability				⬛	⬛	🟫	🟫					🟫
	Confounding Ideas		🟫	🟫									
Interpret Confidence Level	Long-run	Correct							🟥		🟥	🟥	🟥
		Repeated						🟥					
		Procedural	🟫										
	Coverage Probability	Connected									🟥	🟥	
		Prob Width		🟫					🟫				
		Compromise				🟫	🟫	🟫					
	Percentage	Value	🟫	🟫	🟫				🟫				
		Actualized				⬛	⬛	⬛	⬛	⬛			
Confident	Quantifiable Measure	Estimating								🟫	🟫	🟫	🟫
		Process				🟫				🟫	🟫	🟫	
	Sureness/ Belief		🟫	🟫	🟫					🟫	🟫		
	Confusion of Confidence		🟫	🟫	🟫		🟫				🟫		
	Equivalent of Confidence					⬛		⬛	⬛				
Confidence Level	Confidence Coefficient		🟫	🟫									🟫
	Relation to Width			🟫	🟫	🟫	🟫	🟫	🟫				
	Probability and Accuracy		🟫		⬛		🟫	⬛	⬛	⬛			
	Coverage Probability								🟥	🟥	🟥	🟥	

Note. Red indicates productive and correct reasoning. Tan indicates developing reasoning. Black indicates incorrect reasoning.

Figure 9

Classification Diagram of Participants' Conceptualizations of the Confidence Interval



Tiana, Kiara, and Aiden were classified as having *Inconsistent* conceptualizations of the concept of confidence interval. While Aiden shared the conceptualization of the word confident and interpreted the confidence level, he differed from the core pair by using both a correct and incorrect interpretation of the confidence interval. The interpretations the participants in this category appear to be memorized without connection to the concepts motivating them, demonstrating initial conceptualizations of confidence intervals.

Joel, Logan, Jace, and Emma were classified as *Developing* conceptualization of the confidence interval. These four participants were most different on their meanings for the term confident. There does appear to be a progression across this group from Logan to Emma in terms of depth of connections and correctness. Jace and Emma share the most similar conceptualization patterns to the Connecting classification. Emma's conceptualization of the word confident and probability, however, make her conceptualizations incompatible with the core conceptualizations in the Connecting category. The distinction between confident and probability is very important to the understanding of the interpretation of confidence intervals and interpretation of the confidence level, marking the difference between the Connecting and Developing group. In general, the Developing group appears to be constructing simultaneous correct and incorrect conceptualizations of the interpretation of the confidence interval, interpretation of the confidence level, the word confident, and the concept of confidence level. It is evident from this grouping that these students have progressed beyond the surface level understanding of the Inconsistent grouping but have not made the connections and depth of understanding that are present in the Connecting group.

Liam, Brody, Gabe and Diana were classified as having conceptualizations similar to the core responses within the *Connecting* category. These four appeared to hold different conceptualization patterns within the interpretation of the confidence level and their conceptualization of the word confident. These mostly correct conceptualizations mark these participants as different from those of the Developing group, but there was still a sense of incomplete development of a robust conceptualization of confidence intervals within this cluster of participants.

It is interesting to point out that there does not appear to be a relationship between statistical experience and conceptualizations of the interpretation of confidence intervals, the interpretation of confidence levels, the word confident, or the concept of confidence level, with a few exceptions. Tiana, as the only introductory student, was classified as having conceptualizations that align with the Inconsistent category. This would be expected due to her limited exposure to the concept of confidence intervals. Liam was the only graduate student who was classified having conceptualizations that align with the Connecting category. The other two graduate students, Joel and Jace, struggled with conceptualizing the interpretations and concepts beyond a mathematical explanation. This led them to be classified as having Developing conceptualizations. While surprising that graduate students would not be classified as the top category, this struggle with conceptual understanding in graduate students has been documented by other researchers[7]. The intermediate students were split between the Inconsistent and Developing classifications, which is also not necessarily surprising. The two intermediate students in the Inconsistent category discussed a large gap between statistical coursework. This gap could perhaps revert these participants back to conceptualizations similar to introductory students. Finally, most capstone students, or senior statistics students, held conceptualizations classified as Connecting. This is also not surprising because these students have been exposed to many undergraduate statistics courses that typically focus on the application rather than theoretical derivation of statistics. There may be a connection between learning how to communicate an application of statistics, as is common in undergraduate coursework, and developing more connected understanding of statistical concepts, such as confidence intervals.

[7] Interested readers are referred to Green (2010), Green and Blankenship (2014), and Noll (2007, 2011)

CHAPTER 5

DISCUSSION

In this chapter, I summarize and discuss the findings in Chapter 4 with respect to following research questions:

1. What conceptualizations of interpretations of confidence intervals and interpretations of confidence levels do undergraduate and graduate students have?
2. What are some similarities and differences in undergraduate and graduate students' concept images of the concept of confidence intervals exist when they conceptualize interpretations of confidence intervals and interpretations of confidence levels?

This chapter continues with implications and future directions for research and teaching, concluding with the limitations to this study.

Summary of Findings

This first section summarizes the findings of the previous chapter by connecting the findings to relevant literature and identifying new findings.

Conceptualizations of the Interpretations (R1)

From the results presented in Chapter 4 for the first research question of this dissertation study, there are four main findings: 1) confirmation of documented (mis)conceptions in the participants of this dissertation study, 2) new conceptions that were not part of the (mis)conception literature found in the participants, 3) classifications of conceptions for the interpretation of the confidence interval and interpretation for the confidence level, and 4) multiple choice interpretation questions may not accurately represent an individual's conception of the interpretation of a confidence interval and interpretation of the confidence level.

Confirmation of Documented (Mis)conceptions

There were 9 documented (mis)conceptions associated with the interpretation of the confidence interval and the interpretation of the confidence level presented in Table 1. Of these, three were found within the responses participants gave during this dissertation study: 1) probability of the actual value of the parameter being within the interval (Canal & Gutiérrez, 2010; Fidler, 2005), 2) accuracy (Kalinowski et al., 2018), and 3) proportion of the population (Andrade et al., 2014; Andrade & Fernández, 2016; Foster, 2014; Hoekstra et al., 2014). By conceptualizing the interpretation of the confidence level as the "percentage of time" the actual value of the parameter is in the actualized interval, Logan, Joel, Jace, and Emma demonstrated the (mis)conception of *probability of actual value of the parameter within the interval*. Tiana, Joel, and Emma discussed the confidence level as a measure of *accuracy*. Kiara discussed the interpretation of the confidence interval as the range of values within which 95% of the population would be, confirming the presence of the *proportion of the population* (mis)conception.

There were three conceptions demonstrated in this dissertation study that were similar, but not exactly the same, to documented (mis)conceptions presented in Chapter 2: 1) a statement that combined the conceptions proportion of the population (Canal & Gutiérrez, 2010; delMas et al., 2007) and the probability of the actual value of the parameter being within the interval (Andrade et al., 2014; Andrade & Fernández, 2016; Foster, 2014; Hoekstra et al., 2014), 2) a combination of the interval as a range of values for the statistic (Fidler, 2005) and the proportion of sample statistics (delMas et al., 2007; Kalinowski et al., 2018), and 3) the fixed disk conception. Tiana's statement, "the range of where 95% of the true proportion of Alex's (sic) songs in the playlist," for the interpretation of the confidence level was a combination of the *proportion of the population* and *probability of the actual value of the parameter being within the interval*. Diana and Gabe both expressed concern about the statement intended to indicate the confidence interval was the interval for the sample statistics (Interview 2): "Jamie is 93% confident that the mean monthly rent for repeated samples of 100 students at HTSU is between $705 and $793." Both participants appeared to interpret this statement as the proportion of sample statistics that would be within this range, presenting a combination of the documented (mis)conceptions *the interval as a range of values for the statistic* and *the proportion of the sample statistics within the confidence interval*. Confusion over the placement of the action in the interpretation of the confidence interval, documented by Callaert (2007) and Foster (2014), was explored using two statements in Interview 2:

1. Jamie is 93% confident that the actual mean monthly rent for all students at HTSU is between $705 and $793. [action is placed on the interval]
2. Jamie is 93% confident that the actual mean monthly rent for all students at HTSU falls within $705 and $793. [action is placed on the parameter of interest]

All of the participants in this dissertation study selected both statements as correct. The potential problem is that these sentences have been used to classify individuals as having the (mis)conception of *fixed disk*. Most participants in this dissertation study, however, were not classified as conceptualizing the confidence interval as fixed with a changing (random) value of the parameter. Using statements like this that change the placement of the action within the interpretation may not provide the information previous researchers assumed was being assessed.

New Conceptions in Interpretations

The participants in this dissertation study provided interpretation responses that could be categorized into 7 main- and 13 sub-themes (see: Appendix N: Table 24). Two identified conceptualizations have not been discussed in previous (mis)conception literature: 1) using the capture/not capture explanation in interpreting the confidence interval, and 2) discussing the connection between the confidence level and the width of the confidence interval. This subsection summarizes the new conceptualizations found in this dissertation study.

The *Capture/Not Capture* conception contains responses that indicated that once the interval has been calculated, the probability the interval captured the actual value of the parameter is either zero or one, not the confidence level probability. It was a very popular response to probing about the equivalence of the words confident and probability in the interpretation of a confidence interval, with six participants discussing this conception over the course of the two interviews. Of those six participants, two participants (Emma and Jace) still provided statements that were categorized as using the confidence level as a probability incorrectly. This conception may be limited to the research site, as it was commonly used in instruction as the counterexample to the rationale that the interpretation of the confidence interval no longer contains probability.

The second new conceptualization was the use of the connection between the confidence interval and the width of the confidence interval as an interpretation of the confidence level (*Coverage Probability – Probability and Width* and *Coverage*

Probability – Compromise). Participants described the concept of the confidence level to the relationship between the confidence level and the width of the interval, with and without considering the "sweet spot" decision statisticians make when choosing a confidence level. These two conceptualizations are not incorrect ideas but may identify students who have a developing understanding of the interpretation of the confidence level. It is important to understand the relationship between the confidence level and the width of the confidence interval, but it is also important to have a deeper understanding of the confidence level.

Classification of Conceptions of the Interpretations

There were four classifications of similar patterns of conceptualizations within participants' responses for the interpretation of the confidence interval (see: Table 10): 1) correct, 2) not probability, 3) replicate of experiment, and 4) incompatible. Participants with *Correct* responses only discussed the correct interpretation of the confidence interval and discussed the capture/not capture scenario only. Students exhibiting *Not Probability* conceptualizations did not interpret the confidence interval using the traditionally correct interpretation, rather they discussed the long-run interpretation and the capture/not capture scenario. *Replicate of Experiment* contained the responses that discussed the interpretation of the confidence interval using repeated experiments, similar to the long-run interpretation for the confidence level. The final classification was *Incompatible*, which contained responses that were either in conflict with each other (a correct and incorrect interpretation) or confounded interpretations.

There were four classifications of similar patterns of conceptualizations within the responses the participants provided for the interpretation of the confidence level (Table

11): 1) correct, 2) probability and width, 3) parallel, and 4) percentage. The *Correct* classification contained responses that were either the long-run interpretation of the confidence interval or identified the correct connections between the confidence level and the coverage probability. The *Probability and Width* classification contained responses that demonstrate the conception that the confidence level is the probability the actual value of the parameter is in the interval and the conception of relating the confidence level to the width of the confidence interval. The responses within the *Parallel* classification indicated that these individuals held conceptualizations for the interpretation of the confidence level as: 1) the correct long-run interpretation of the confidence level, 2) the confidence level as related to the width of the confidence interval, and 3) the confidence level as related to the probability the actual value of the parameter will be within the actualized interval. The final classification category is *Percentage*, which contained responses that the confidence level represents the probability that a value (parameter or population value) is within the interval.

Closed-response versus Open-response Interpretations

Most of the conceptions documented in the previous two sub-sections were found when participants responded to open-middle questions from Interview 1. On Interview 2, the participants were given statements that mimicked documented (mis)conceptions from the literature and presented in ways that were similar to closed-response questions (multiple statements provided on one sheet of electronic paper, for statements see: Table 8, Table 9, and Table 12). Based on the categorizations of responses from Interview 1, it appeared that all of the participants would not share the same beliefs about the correctness and incorrectness of the provided statements. The analysis of the statements from Interview 2, however, indicated that, despite have four different classifications of conceptualization patterns among the participants, most participants agreed on the correctness or incorrectness of the provided statements. Particularly troubling are participants who could select correct interpretations on Interview 2, but held no deeper understanding of the statement. For instance, Kiara could choose the correct interpretation of the confidence interval and state the correct interpretation of the confidence interval. She stated explicitly, however, that she was unable to explain what the confidence level or the word confident meant. On closed-form assessments that provide multiple choice, true/false, and/or multiple select questions (such as CAOS, LOTUS, etc.), individuals like Kiara may be able to demonstrate proficient understanding of confidence intervals because they are able to select the correct interpretation but have not internalized more than a procedural memorization of the interpretation. (Mis)conception studies presented in the literature review such as Hoekstra et al. (2012), Crooks et al. (2019), García-Pérez and Alcalá-Quintana (2016), and Canal and Ruiz (2015), presented their results based on closed-form assessments. From the results of this dissertation study, it may be the case that these closed-form assessments are over-estimating the number of individuals with correct knowledge of confidence intervals. Furthermore, the results presented in Chapter 4: *Similarities and Differences in Concept Images (R2)* (summarized in the next section) demonstrate that these statements do not provide the full picture of the knowledge and understanding the participants have about the interpretation of the confidence interval and the interpretation of the confidence level. The many different conceptualizations of confident and confidence level indicate that the method of closed-form assessments may not provide the necessary information to properly assess understanding of the interpretations.

Similarities and Differences of Concept Image (R2)

As participants discussed their understanding of the interpretation of a confidence interval and the interpretation of the confidence level, it became evident there were many similarities and differences among the dimensions of the participants' concept images that could explain the differences in interpretations. Five dimensions that appeared to contain similarities and differences across the participants are: 1) the meaning of the word confident, 2) the concept of confidence level, 3) the visualization of the confidence interval, 4) the random process behind the confidence interval (The Jamie's Colleague Task), and 5) probability. The first two were explored in this dissertation study. With little research focused on these dimensions, the conceptualizations of the word confident and the concept of confidence level are new findings and are summarized below, with relevant connections to literature. These last three dimensions are addressed in the *Future Directions* section to follow.

Confident

There is little literature in the statistics education research on how individuals define the word confident, with the exception of Kaplan et al. (2010). Kaplan et al. were focused on the lexical ambiguity of the word confident and found five general themes to their student responses. They found that the majority of their participants defined the word confident in a way that implied a level of surety or certainty. The categories found in this dissertation study, however, focused on the use of the word confident in the interpretation of the confidence interval rather than the colloquial versus statistical definition of the word confident, as was the case in Kaplan et al. Four conceptualization categories were created from the definitions and implications the participants of this dissertation study provided: 1) Quantifiable Measure, 2) Sureness/Belief, 3) Confusion about Confident, Chance, and Probability, and 4) Equivalence of Confident, Chance, and Probability. This section discusses the findings from this dissertation study about the categorization of the participants' responses.

The *Quantifiable Measure* category of conceptualizations of the word confident contained statements that implied the word confident to mean that the confidence interval was "doing what it was supposed to" or was referring to the statistical procedure being conducted. The *Sureness/Belief* conceptualization, which overlapped with categories found by Kaplan et al., defined the word confident as equivalent to sure, certainty, and belief, all of which can be used to express ideas of formal or informal probabilities. Further research will be needed to explore whether using sure or belief as equivalent to confident affects the interpretation of the confidence interval and/or the interpretation of the confidence level. The *Confusion about Confidence, Chance, and Probability* conceptualization contained responses that either expressed initial confusion or responses that were contradictory over the course of the two interviews about the equivalence of the words confidence, chance and probability. The responses in the *Equivalence of Confidence, Chance, and Probability* category stated explicitly that the three words were equivalent.

In general, there were four classifications for similar patterns of conceptualizations of the word confident: 1) Quantifiable Measure, 2) Belief, 3) Towards Probability, and 4) Probability. *Quantifiable Measure* only consisted of the conceptualization that confidence was a way to quantify the confidence interval. *Belief* is the conceptualization of the word confident as a quantifiable measure of sureness/belief the actual value of the parameter will be between the upper and lower bound of the confidence interval. *Toward Probability* consists of conceptualizations that imply the equivalence of confident to sureness and/or belief but also expressed contradictory statements about whether confident and chance were equivalent. The final classification was *Probability*, which only contained the conceptualization that confidence, probability, chance, and sure were equivalent statements. The implication and future directions for this dimension of the concept image will be discussed in the next two sections.

Confidence Level

The concept of the confidence level appears to be another dimension of the concept image along which similarities and differences existed. From the literature, the only documented conception that is similar to identified conceptualizations of confidence level is the equivalence of the word confident to accuracy (Kaplan et al., 2014). The participants demonstrated various depths of understanding with respect to the confidence level. As a theoretically complicated concept, it is not surprising that students recruited from lower levels of statistics courses did not demonstrate the same depth of understanding as more advanced statistics students.

There were four categorizations to the conceptualizations of the confidence level (see: Table 19): 1) confidence coefficient, 2) relation to width, 3) probability and accuracy, and 4) coverage probability. *Confidence coefficient* was the conceptualization that was procedural in nature. Participants that held this conceptualization discussed the confidence level as a way to calculate the confidence interval rather than a deeply connected statistics concept. The *Relation to Width* conceptualization extended the confidence coefficient conception to indicate how the confidence level effects the margin of error and the width of the confidence interval. *Probability and Accuracy* conceptualization identifies the confidence level as either the probability the actual value of the parameter is between the upper and lower bounds of the confidence interval or the accuracy that the confidence interval captured the actual value of the parameter. The final category of conceptualizations, *Coverage Probability*, contained four parts, of which a participant needed to discuss at least one to be part of this category: 1) the confidence level is related to the proportion of statistics within the middle region of the sampling distribution centered at a value of the unknown parameter (see Figure 7a), 2) explaining the choice of the margin of error as half the width of the confidence level% region of the sampling distribution (Figure 7b), 3) connecting the proportion of statistics within the middle region of the sampling distribution to the long-run interpretation of the confidence level, and 4) relating the long-run interpretation with the probability of randomly selecting a sample from the population. As with the meaning of the word confidence, the concept of confidence level appears to be important in developing correct and deeply connected understanding of confidence intervals. This dissertation study was not able to fully explore its importance and is addressed in the *Future Directions* section.

There were three classifications of conceptualization themes: 1) coverage probability, 2) developing, and 3) surface level. The *Coverage Probability* classification only contained conceptualizations that were part of the four-component coverage probability conception, described above. Those in the *Developing* classification held conceptions somewhere between the two participants with well-connected ideas about the coverage probability and the surface level conceptions. The *Surface Level* classification contains participants who strictly conceptualize the confidence level as it related to the calculation of the confidence interval. The implications of the depth of understanding of the confidence level and coverage probability will be discussed in the next two sections.

Implications and Future Directions

This section discusses the implications for research and teaching that have arisen as a result of this dissertation study. Five implications for future research and three implications for teaching are discussed. As a generative research study, it has exposed more questions about the knowledge required to deeply understand confidence intervals. Therefore, the implications of this dissertation study for future research are expansive.

Implications for Research

Implications for future research are: 1) pursuing the interconnectedness of the interpretations of the confidence interval, the interpretation of the confidence level, confident, and confidence level, 2) analyzing the three additional dimensions of similarities and differences among participant's concept images, 3) testing the hypothesized concept image and developmental cloud for confidence intervals, 4) expanding the understanding of the *Developing* classification of the concept of confidence intervals, 5) exploring the conceptualizations presented in this dissertation to identify productive or non-productive reasoning and/or are a result of epistemological or didactical obstacles.

Interconnectedness of the Interpretations and Curricular Concepts

The results presented in this dissertation, demonstrate the need to investigate the interconnectedness of the interpretation of the confidence interval and the interpretation of the confidence level. From the (mis)conception literature presented in Chapter 2, it appeared that the interpretation of the confidence interval and the interpretation of the confidence level were disjoint concepts. The two analyzed dimensions of the concept image of confidence intervals provide evidence that individual’s conceptions of the interpretations of the confidence interval and the interpretation of the confidence level are not easily separated. An individual’s personal concept definitions of confident and confidence level appear to be connected to whether an individual conceptualizes the interpretations in a probabilistic way. This dissertation study stopped short of being able to fully identify the implications of individual’s personal concept definitions of confidence and confidence level on the interpretation of the confidence interval and interpretation of the confidence level. Now that these two dimensions have been exposed as potentially important aspects of an individual’s concept image of confidence intervals, further research studies should be conducted to identify the relationship between these two dimensions. Further, by focusing a research study on one or both of these dimensions, researchers may be able to identify the effect these personal concept definitions have on the interpretations more concisely than in the findings for this dissertation study.

Additional Dimensions to be Analyzed

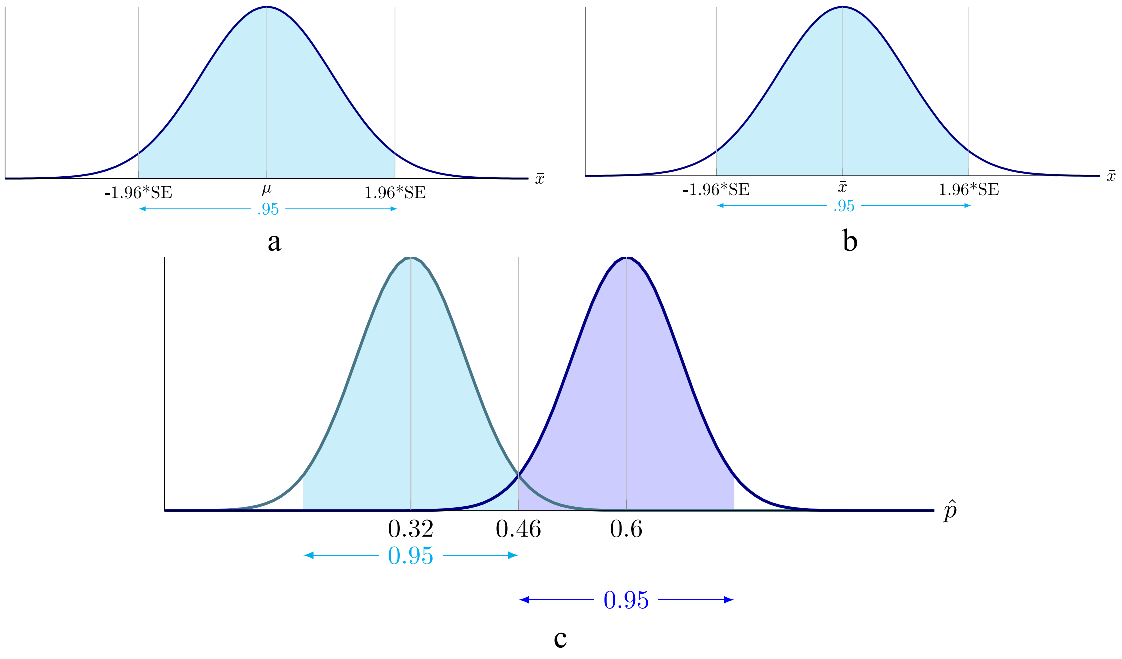
Three additional dimensions were identified as areas of similarities and differences within the concept image of the confidence interval: 1) visualization of the confidence interval, 2) random process (The Jamie’s Colleague Task), and 3) conceptualization of probability. There were three distinct categories of *visualization* of the confidence interval that were described by the participants of this dissertation study:

1) a normal distribution centered at the statistic (Figure 10b), 2) a hypothesis test (similar to Figure 10a, but centered at the null hypothesis value), and 3) alternative (similar to Figure 10c). There appears to be a connection between the visualization of the confidence interval using Figure 10b and fully understanding the implications of the coverage probability on the interpretation of the confidence interval and the interpretation of the confidence level. The interested reader is referred to Emma's and Brody's summaries found in Appendix L to see initial analysis of the effect the difference in visualization has on the conceptualization of confidence intervals. The Capture/Not Capture conception found during Interview 1 lead to the creation of The Jamie's Colleague Task on Interview 2 (Appendix I: Task 2), which was used to explore whether or not participants understood the explanation they were providing in the first interview. Initial analysis of this task indicates that participants struggled with identifying the difference between the interval estimator and the interval estimate. In particular, participants struggled with the idea that prior to collecting data there was probability of the random interval capturing the actual value of the parameter, but either a probability of 0 or 1 that the actualized interval captured the actual value of the parameter. Understanding the existence of probability prior to collecting the data, appears to be connected to understanding the ideas of random process and coverage probability but requires further analysis. Finally, three participants presented potentially problematic views of probability. Most noticeably was Kiara's statement indicating that 95% confidence implied "more likely than not," which indicates a possible conception that any probability greater than 50% is most likely to happen. Further research will need to be done to see if particular conceptualizations of probability

have an effect on the interpretation of the confidence interval and the interpretation of the confidence level.

Figure 10

Example of Visualizations of Confidence Intervals

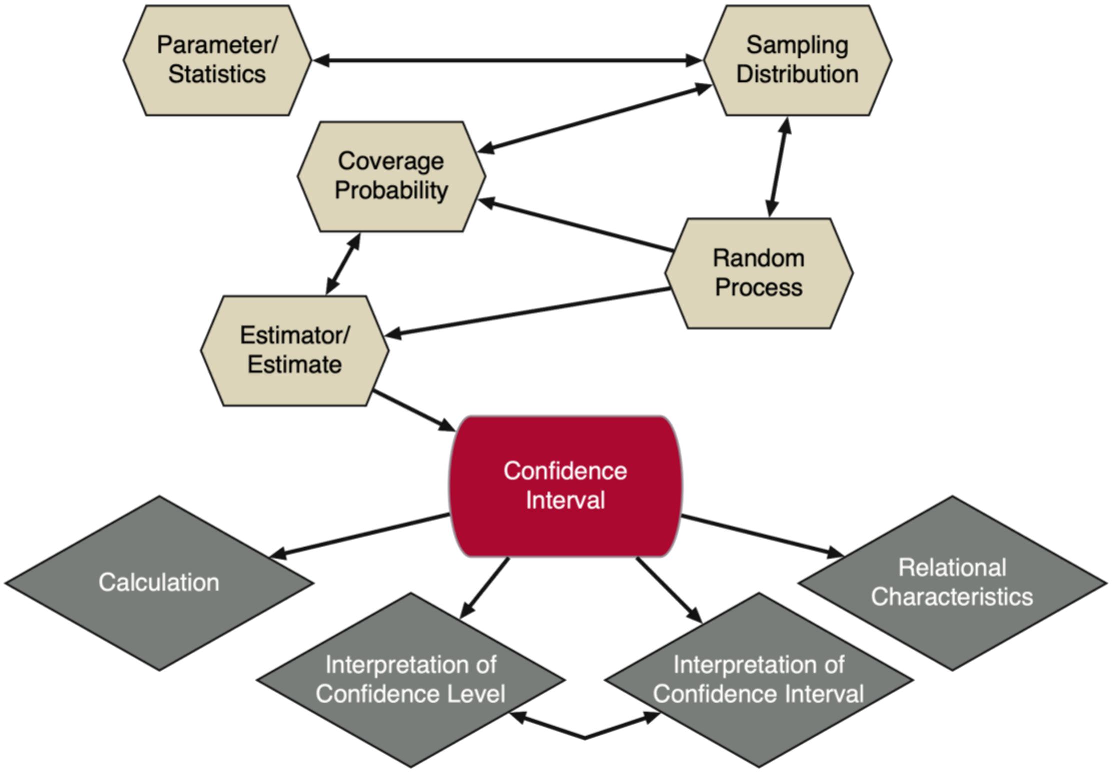


Further Development of the Concept Image for Confidence Intervals

Another implication for future research consists of refining the proposed learning clouds in the concept image of confidence intervals. The results from this dissertation study exposed the need to refine the current concept image of confidence intervals (found in developmental cloud form in Figure 3, Figure 4, and Figure 5). Figure 11, projecting a new hypothetical developmental cloud, demonstrates a possible refinement. The implication of the interconnectedness of the interpretations of the confidence interval and confidence level is that there may not need to be individual concept images for the two interpretations. Instead, the learning clouds may need to be coordinated prior to being used in the interpretation of the confidence interval and the interpretation of the confidence level. The identified dimensions of confident and confidence level suggest the importance of understanding the ideas within the Coverage Probability, Random Process, and Estimator/Estimate learning clouds. It may be necessary, however, to redefine learning clouds to focus on the aspects of similarities and differences that were identified in this dissertation study. Rather than having learning clouds that were hypothesized based on potential statistical concepts needed to understand confidence intervals robustly, learning clouds may need to be defined base on dimensions such as visualization, random process, confidence level, and confident. Further research into this new concept image and exploration into the learning clouds will help to confirm this new hypothesized developmental cloud.

Figure 11

New Hypothesized Developmental Cloud for the Concept of Confidence Intervals



Classification of Conceptualizations of Confidence Intervals

There were three classifications for the conceptualization of confidence intervals presented in this dissertation study: 1) Inconsistent, 2) Developing, and 3) Connecting. Each category had a pair of participants whose conceptualizations were used to develop the core ideas of these categories. The Developing classification represent a cluster of conceptualization patterns that require further exploration. The Inconsistent and Connecting conceptualizations appear to demonstrate possible initial or surface level understanding and deep understanding of confidence intervals, respectively. Future research could help understand the productive pathway to robust understanding of confidence intervals. The next two sub-sections expand the importance of understanding these conceptualizations better.

Conceptions as Productive and Non-Productive Reasoning and/or Obstacles

The final two implications for research consist of further exploring and identifying the productivity of conceptualizations identified in this dissertation study with respect to the interpretation of the confidence interval and the interpretation of the confidence level. As indicated in the previous section, there appears to be two solid groupings of conceptualizations that indicate low depth of understanding and high depth of understanding. This begs the question of which conceptions have the possibility to be productive paths to deeper understanding? Which conceptions could derail productive reasoning and become non-productive paths for deep understanding? Further research will be required to answer these questions. Additionally, in Chapter 2, I proposed that some instruction introduces didactical obstacles that could be potentially avoided with improved instructional methods. Future research could help identify if any of the conceptions that have been reported in this dissertation study have been instilled as part of the didactical transposition of the concept of confidence intervals. Conceptions such as the Confidence Coefficient and Relation to Width may be present in individuals because introductory instruction focuses on the procedure of finding confidence coefficients and margin of error rather than grounding the confidence level in the curricular concepts that motivate it. Further research will be needed to investigate this hypothesis. The next section does, however, propose a new method of teaching confidence intervals that grounds the confidence level in the components of the conceptualization Coverage Probability of the concept of Confidence Level.

Implications for Teaching

There are three recommendations for teaching from the findings of this dissertation study: 1) the use of the capture/not capture scenario in instruction, 2) the need to connect the confidence level with the ideas of the coverage probability, and 3) a proposed method for introducing confidence intervals. The first implication is a cautious recommendation. There appears to be some benefit to the Capture/Not Capture conceptualization, but only if the individual has also assimilated a non-probability conception of the confidence level. This explanation is a helpful counterexample to the conception that there is probability of the actual value of the parameter being within the upper and lower bounds of the confidence interval. It needs to be grounded, however, in the understanding that probability is associated with a random variable, and thus the estimator, rather than the actualized interval. The second recommendation is connected to the depth of knowledge of the coverage probability and confidence level that was demonstrated by the participants in this study. Quite a few participants held no conception of the confidence level beyond its use in calculating the confidence coefficient or the width of the interval. Thus, those participants, without understanding the confidence level and the broader concept of coverage probability, were unable to interpret the confidence level, leaving the confidence level open to a probabilistic interpretation. The third recommendation attempts to implement the prior two recommendations. The results of this dissertation study demonstrate the importance of understanding the coverage probability to the interpretation of the confidence interval, the interpretation of the confidence level, the word confident, and the confidence level. This is potentially problematic in an introductory course, but the results from this dissertation study seem to suggest that it is important in developing sound conceptualizations of the interpretations of confidence intervals and the interpretation of the confidence level. Interview 3 (see: Appendix J: Task 2) proposed a task that discussed the connection among the four components identified in the Coverage Probability category of the concept of confidence level (see Table 13 for a list of the components). Further research will be needed to focus on identifying if the task aids in productive formation of the conceptualization of the confidence interval that helps students understand the interpretation of the confidence interval and the interpretation of the confidence level better.

Limitations

This dissertation study focused on the conceptualizations of the interpretation of confidence intervals and interpretation of confidence level using a qualitative study to better understand the cognitive structure of these conceptualizations. Qualitative analysis depends heavily on the researchers' subjectivity and perspective. Therefore, unlike quantitative analysis which produces similar results from the same core data and analysis process, qualitative analysis may differ from researcher to researcher. There are inherent biases within a qualitative study that may not be present in quantitative studies. For instance, the small sample size and recruitment procedures do not allow these findings to be generalized to a broader population. The participants of this study self-selected to participate and continue in this study and may not represent the average student because of their internal motivation to participant in a research study.

The interviews for this dissertation study took place over the span of three months. During this time, all but Tiana were enrolled in statistics classes. The continuation of instruction could have changed the conceptions that the participants held from interview to interview. Kiara and Diana both recalled more information during the second and third interview than in the first interview, confirming this limitation. In future studies, I would add additional time and limit the number of participants in the study to better tailor questions and protocols to each participant allowing better understanding of a person's concept image. This lack of individualization may have limited the findings in this dissertation study and may have led to less fine-grain findings. This dissertation study design incorporated three interviews per participant. Additional interviews would have allowed me to capture more data and gather more information about each participant's conceptualizations. Finally, the method of data collection is not a natural situation for either participant or interviewer. Being recorded and asked to describe their thinking could have impacted the knowledge that each participant demonstrated. While I attempted to make each interview conversational and distract the participants from the video cameras, it is possible that the situation biased the participants' data.

REFERENCES

Agresti, A., Franklin, C., & Klingenberg, B. (2017). *Statistics: The art and science of learning from data* (4th ed.). Pearson.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed). American Psychological Association.

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(Suppl 1.), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Andrade, L., & Fernández, F. (2016). Interpretation of confidence interval facing the conflict. *Universal Journal of Educational Research*, *4*(12), 2687–2700. https://doi.org/10.13189/ujer.2016.041201

Andrade, L., Fernández, F., & Álvarez, I. (2014). Fostering changes in confidence intervals interpretation. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*. http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C187_ALVAREZ.pdf?1405041858

ASA GAISE College Report Revision Committee. (2017). *Guidelines for Assessment and Instruction in Statistics Education (GAISE): College Report 2016*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

Bell, A., & Burkhardt, H. (2002). *Domain frameworks in mathematics and problem solving*. Annual Meeting of the American Education Research Association, New Orleans, LA. https://www.mathshell.com/papers/pdf/domains.pdf

Bertie, A., & Farrington, P. (2003). Teaching confidence intervals with Java applets. *Teaching Statistics*, *25*(3), 70–74. https://doi.org/10.1111/1467-9639.00134

Brousseau, G. (1997). *Theory of didactical situations in mathematics: Didactique des mathématiques, 1970–1990* (N. Balacheff, R. Sutherland, & V. Warfield, Eds.; Vol. 19). Kluwer Academic Publishers.

Callaert, H. (2007). Understanding confidence intervals. *Proceedings of the 5th Congress of the European Society for Research in Mathematics Education*, 692–701. http://www.mathematik.uni-dortmund.de/~erme/CERME5b/WG5.pdf

Canal, G. Y., & Gutiérrez, R. B. (2010). The confidence intervals: A difficult matter, even for experts. *Proceedings of Eighth International Conference on Teach Statistics (ICOTS8)*. https://iase-web.org/documents/papers/icots8/ICOTS8_C143_CANAL.pdf?1402524973

Canal, G. Y., & Ruiz, L. R. (2015). On the meaning that teachers in training will give to the accuracy of a confidence interval and its relationship with the sample size and level of confidence. *Proceedings of the 2015 Satellite Conference of the International Association for Statistical Education (IASE)*. https://iase-web.org/documents/papers/sat2015/IASE2015%20Satellite%2082_YEZ.pdf?1438922683

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed). Thomson Learning.

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. https://doi.org/10.1016/j.edurev.2007.04.001

Chevallard, Y., & Bosch, M. (2014). Didactic transposition in mathematics education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 170–174). Springer Netherlands. https://doi.org/10.1007/978-94-007-4978-8_48

Chihara, L., & Hesterberg, T. (2011). *Mathematical statistics with resampling and R*. Wiley.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 326–353). Routledge.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823. https://doi.org/10.2307/2975286

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00026

Crooks, N. (2014). *Does comparison promote gains in conceptual knowledge? The case of learning about confidence intervals* [Doctoral dissertation, University of Wisconsin-Madison]. ProQuest.

Crooks, N., Bartel, A., & Alibali, M. W. (2019). Conceptual knowledge of confidence intervals in psychology undergraduate and graduate students. *Statistics Education Research Journal*, *18*(1), 46–62. https://iase-web.org/documents/SERJ/SERJ18(1)_Crooks.pdf?1558844352

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics*, *29*(3), 89–93. https://doi.org/10.1111/j.1467-9639.2007.00267.x

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–180. https://doi.org/10.1037/0003-066X.60.2.170

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*(3), 217–227. https://doi.org/10.1037/1082-989X.11.3.217

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*(4), 299–311. https://doi.org/10.1207/s15328031us0304_5

Cumming, J., Miller, C., & Pfannkuch, M. (2014). Using bootstrap dynamic visualizations in teaching. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_3D1_CUMMING.pdf?1405041608

De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2018). *Intro stats* (5th ed). Pearson.

delMas, R. C., Garfield, J., Ooms, A., & Chance, B. L. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58. https://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf?1402525007

diSessa, A. A. (2006). A history of conceptual change research: Threads and fault lines. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 265–281). Cambridge University Press.

diSessa, A. A. (2007). An interactional analysis of clinical interviewing. *Cognition and Instruction*, *25*(4), 523–565. https://doi.org/10.1080/07370000701632413

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology* [Doctoral dissertation, The University of Melbourne]. https://fionaresearch.files.wordpress.com/2013/06/fidler-phd-2006.pdf

Fidler, F. (2006). Should psychology abandon p values and teach cis instead? Evidence-based reforms in statistics education. *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS7)*. https://iase-web.org/documents/papers/icots7/5E4_FIDL.pdf?1402524965

Foster, C. (2014). Confidence trick: The interpretation of confidence intervals. *Canadian Journal of Science, Mathematics and Technology Education*, *14*(1), 23–34. https://doi.org/10.1080/14926156.2014.874615

Fricker, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in basic and applied social psychology after their p -value ban. *The*

American Statistician, *73*(Suppl. 1), 374–384. https://doi.org/10.1080/00031305.2018.1537892

García-Pérez, M. A., & Alcalá-Quintana, R. (2016). The interpretation of scholars' interpretations of confidence intervals: Criticism, replication, and extension of Hoekstra et al. (2014). *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01042

Gilliland, D., & Melfi, V. (2010). A note on confidence interval estimation and margin of error. *Journal of Statistics Education*, *18*(1). https://doi.org/10.1080/10691898.2010.11889474

Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education*. Lawrence Erlbaum Associates.

Goodman, S. N. (2019). Why is getting rid of p-values so hard? Musings on science and statistics. *The American Statistician*, *73*(Suppl. 1), 26–30. https://doi.org/10.1080/00031305.2018.1558111

Gordon, S. P., & Gordon, F. S. (2020). Visualizing and Understanding Hypothesis Testing Using Dynamic Software. *PRIMUS*, *30*(2), 172–190. https://doi.org/10.1080/10511970.2018.1534295

Grant, T. S., & Nathan, M. J. (2008). *Students' conceptual metaphors influence their statistical reasoning about confidence intervals* (WCER Working Paper No. 2008-5). University of Wisconsin-Madison. https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2008_05.pdf

Green, J. L. (2010). Teaching highs and lows: Exploring university teaching assistants' experiences. *Statistics Education Research Journal*, *9*(2), 108–122. https://iase-web.org/documents/SERJ/SERJ9(2)_Green.pdf?1402525009

Green, J. L., & Blankenship, E. E. (2014). Beyond calculations: Fostering conceptual understanding in statistics graduate teaching assistants. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_3A3_GREEN.pdf?1405041598

Hagtvedt, R., Jones, G. T., & Jones, K. (2007). Pedagogical simulation of sampling distributions and the central limit theorem. *Teaching Statistics*, *29*(3), 94–97. https://doi.org/10.1111/j.1467-9639.2007.00270.x

Hagtvedt, R., Jones, G. T., & Jones, K. (2008). Teaching confidence intervals using simulation. *Teaching Statistics*, *30*(2), 53–56. https://doi.org/10.1111/j.1467-9639.2008.00308.x

Henriques, A. (2016). Students' difficulties in understanding of confidence intervals. In D. Ben-Zvi & K. Makar (Eds.), *The Teaching and Learning of Statistics* (pp. 129–138). Springer International Publishing. https://doi.org/10.1007/978-3-319-23470-0_18

Hoekstra, R., Johnson, A., & Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement*, *72*(6), 1039–1052. https://doi.org/10.1177/0013164412450297

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hoekstra, R., Morey, R. D., & Wagenmakers, E.-J. (2018). Improving the interpretation of confidence and credible intervals. *Proceedings of the International Conference on Teaching Statistics (ICOTS10)*. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_8A2.pdf?1531364291

Hunting, R. P. (1997). Clinical interview methods in mathematics education research and practice. *Journal of Mathematical Behavior*, *16*(2), 145–165. https://doi.org/10.1016/S0732-3123(97)90023-7

Inzunza, S. (2018). Design and evaluation of a hypothetical learning trajectory to confidence intervals based on simulation and real data. *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10)*. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_9H1.pdf?1531364301

Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. *Preceedings of the Eighth International Conference on Teaching Statistics (ICOTS8)*. https://iase-web.org/documents/papers/icots8/ICOTS8_C104_KALINOWSKI.pdf?1402524972

Kalinowski, P., Lai, J., & Cumming, G. (2018). A cross-sectional analysis of students' intuitions when interpreting cis. *Frontiers in Psychology*, *9*, 1–19. https://doi.org/10.3389/fpsyg.2018.00112

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2010). Lexical ambiguity in statistics: How students use and define the words: Association, average, confidence, random and spread. *Journal of Statistics Education*, *18*(2), 1–6. https://doi.org/10.1080/10691898.2010.11889491

Kaplan, J. J., Roland, K. E., Woodard, V. L., & Woodard, R. D. (2018). *Stat 2000 introductory statistics laboratory manual: StatCrunch Edition*. Macmillan Learning Curriculum Solutions.

Koklu, O. (2017). *Undergraduate students' informal notions of variability* [Doctoral dissertation, The University of Georgia]. https://iase-web.org/documents/dissertations/17.OguzKoklu.Dissertation.pdf

Krall, J. (2016, December 20). *The biggest stats lesson of 2016*. Sense About Science USA. https://senseaboutscienceusa.org/biggest-stats-lesson-2016/

Larson, R., & Farber, E. (2019). *Elementary statistics: Picturing the world* (7th ed.). Pearson Education, Inc.

Liu, Y. (2005). *Teachers' understandings of probability and statistical inference and their implications for professional development* [Doctoral dissertation, Vanderbilt University, Vanderbilt University]. ProQuest.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the power of data*. Wiley.

Matthews, R. A. J. (2019). Moving towards the post p < 0.05 era via the analysis of credibility. *The American Statistician*, *73*(Suppl. 1), 202–212. https://doi.org/10.1080/00031305.2018.1543136

McClave, J. T., & Sincich, T. (2017). *A first course in statistics* (12th ed.). Pearson.

Miller, J., & Ulrich, R. (2016). Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). *Psychonomic Bulletin & Review*, *23*(1), 124–130. https://doi.org/10.3758/s13423-015-0859-7

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, *10*(1). https://doi.org/10.1080/10691898.2002.11910548

Moore, D. S., & Notz, W. (2009). *Statistics: Concepts and controversies* (7th ed.). Freeman.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Noll, J. (2007). *Graduate teaching assistants' statistical knowledge for teaching* [Doctoral dissertation, Portland State University]. https://iase-web.org/documents/dissertations/07.Noll.Dissertation.pdf

Noll, J. (2011). Graduate teaching assistants' statistical content knowledge of sampling. *Statistics Education Research Journal*, *10*(2), 48–74. https://iase-web.org/documents/SERJ/SERJ10(2)_Noll.pdf?1402525003

Peck, R. (2014). *Statistics: Learning from data*. Brooks/Cole.

Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *The Journal of Mathematical Behavior*, *22*(4), 405–435. https://doi.org/10.1016/j.jmathb.2003.09.002

Rossman, A. J., & Chance, B. L. (2012). *Workshop statistics: Discovery with data* (4th ed). Wiley.

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. https://doi.org/10.1207/s15327809jls0302_1

Steel, E. A., Liermann, M., & Guttorp, P. (2019). Beyond calculations: A course in statistical thinking. *The American Statistician*, *73*(Suppl. 1), 392–401. https://doi.org/10.1080/00031305.2018.1505657

Sullivan, M. (2013). *Statistics: Informed decisions using data* (4th ed). Pearson.

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, *12*(2), 151–169. https://doi.org/10.1007/BF00305619

Thompson, P. W. (2008). Conceptual analysis of mathematical ideas: Some spadework at the foundation of mathematics education. *Proceedings of the Joint Meeting of PME 32 and PME-NA XXX*. http://pat-thompson.net/PDFversions/2008ConceptualAnalysis.pdf

Thompson, P. W., Carlson, M. P., Byerley, C., & Hatfield, N. (2014). Schemes for thinking with magnitudes: An hypothesis about foundational reasoning abilities in algebra. In K. C. Moore, L. P. Steffe, & L. L. Hatfield (Eds.), *Epistemic algebra students: Emerging models of students' algebraic knowing* (Vol. 4, pp. 1–24). http://pat-thompson.net/PDFversions/2013MagsInAlg.pdf

Thompson, P. W., & Liu, Y. (2005). Understandings of margin of error. In S. Wilson (Ed.), *Proceedings of the Twenty-seventh Annual Meeting of the International*

Group for the Psychology of Mathematics Education. http://pat-thompson.net/PDFversions/2005PMENA%20MOE.pdf

Tintle, N., Chance, B. L., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2015). Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. *The American Statistician*, *69*(4), 362–370. https://doi.org/10.1080/00031305.2015.1081619

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2015). *Introduction to statistical investigations* (Preliminary ed). Wiley.

Triola, M. F., & Iossi, L. (2018). *Elementary statistics* (13th edition). Pearson.

Utts, J. M. (2005). *Seeing through statistics* (3rd ed). Thomson, Brooks/Cole.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2002). *Mathematical statistics with applications* (6th ed). Duxbury.

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. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019a). Moving to a World Beyond "p < 0.05." *The American Statistician*, *73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (Eds.). (2019b). Statistical Inference in the 21st Century: A World Beyond p < 0.05 [Special Issue]. *The American Statistician*, *73*(Suppl. 1).

Weiss, N. A., & Weiss, C. A. (2016). *Introductory statistics* (10th edition). Pearson.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Williams, I. J., & Williams, K. K. (2018). Using an R shiny to enhance the learning experience of confidence intervals: Using an R shiny to enhance the learning experience of confidence intervals. *Teaching Statistics*, *40*(1), 24–28. https://doi.org/10.1111/test.12145

Yeo, J. B. W. (2017). Development of a Framework to Characterise the Openness of Mathematical Tasks. *International Journal of Science and Mathematics Education*, *15*(1), 175–191. https://doi.org/10.1007/s10763-015-9675-9

Zieffler, A., & Catalysts for Change. (2018). *Statistical thinking: A simulation approach to modeling uncertainty* (4.1th ed.). Catalyst Press. http://zief0002.github.io/statistical-thinking/

APPENDICES

APPENDIX A

TABLE OF DEFINITIONS

- *Actualized or realized*: a descriptor to make the distinction between data that have been gathered and a theoretical sample that has yet to be collected. An actualized/ realized sample is one that has been collected.
- *Actualized variable*: a variable calculated from collected data.
- *Categorical variable*: a characteristic attribute.
- C*oncept image*: "the total cognitive structure that is associated with the concept, which includes all the mental pictures and associated properties and processes. It is built up over the years through experiences of all kinds, changing as the individual meets new stimuli and matures" (Tall & Vinner, 1981, p. 152).
- *Conception*: as a word to maintain a less negative connotation and could "identify and relate factors that students use[d] to explain intriguing or problematic phenomena" (Smith et al., 1994, p. 119).
- *Confidence coefficient*: The infimum, the greatest lower bound of a set, of the coverage probabilities. Formally: $inf_{\theta} P_{\theta}(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$.
- *Confidence level*: the proportion of statistics from the middle of the true sampling distribution for the statistic generated from samples of size n centered at the actual value of the parameter, i.e. middle confidence level % of the sampling distribution.

- *Confidence interval estimate*: a form of an interval estimate, with a coinciding confidence level, derived from a confidence interval estimator.
- *Confidence interval estimator*: a form of an interval estimator that had been derived based on the desired coverage probability.
- *Confident*: a word to refer to the difference between the calculated interval, which no longer has probability of capturing the actual value of the parameter, and the random interval, which has probability associated with the random process through which the interval was constructed.
- *Coverage probability:* This was the probability that the random *interval estimator* covered the actual value of the parameter θ. Formally: $P_\theta(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$ or $P([L(\boldsymbol{x}), U(\boldsymbol{x})]|\theta)$.
- *Datum*: an attribute of a unit of interest.
- *Definitional knowledge*: the knowledge that an actualized confidence interval is an estimate for an unknown value of the population parameter and is part of the inferential family (Fidler, 2005).
- *Developmental cloud*: "an ensemble of meanings and ways of thinking … entail[ing] some common ways of thinking while at the same time involving ways of thinking that are unique to [the individual]" (Thompson et al., 2014, p. 14). The developmental cloud is a way of visualizing of the coordination of many different meanings and ways of thinking to understand the larger concept of confidence intervals.
- *Deviation*: the difference each value of a variable was from the mean of the variable.

- *Didactical transposition*: transforming contextualized subject knowing into curricular knowledge.
- *Distribution*: a representation of a variable.
- *Distribution of a population*: the graphical representation of all values of the variable of interest from a population.
- *Distribution of a sample*: the graphical representation of all values of the variable of interest from a sample.
- *Estimate*: the calculated value of an estimator calculated using an actualized collection of data.
- *Estimator*: the function of a random variable (i.e. point estimator, interval estimator).
- *Empirical probability distribution*: the actualized representation of the long-run relative frequency of a random variable based on a collection of outcomes from a random experiment.
- *Evoked concept image*: "a portion of the concept image which is activated at a particular time" (Tall & Vinner, 1981, p. 152).
- *Formal concept definition*: the formal mathematical (or statistical) definition of a curricular concept.
- *Fundamental Confidence Fallacy* described as :"If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observed interval contains the true value is also X%; or, alternatively, we can have X% confidence that the observed interval contains the true value" (Morey et al., 2016, p. 104).

- *Independent*: the idea that the result(s) from a previous trial had no effect on the results from the current trial.
- *Independent random variable*: a random variable that was produced from repeated independent trials of a random experiment.
- *Interpretation of a Confidence Interval*: We are CL% confident that the actual value of the parameter is between the upper and lower bounds of the interval.
- *Interpretation of a Confidence Level*: Approximately confidence level% of all possible samples of size n from a population would produce confidence level% confidence intervals that capture the actual value of the parameter.
- *Interval estimate*: An interval estimate was a realized interval that was created from an interval estimator with a realized random sample that can be used to estimate a value of the unknown parameter, θ.
- *Interval estimator*: An interval estimator was an interval that was created from two functions of the random variable such that a random interval was created. Formally, given any pair of functions $L(x_1, \dots, x_n)$ and $U(x_1, \dots, x_n)$ of a sample that satisfy $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$, then the random interval $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ was called an *interval estimator.*
- *Learning clouds*: the concept that Thompson et al. (2014) created to visually describe the idea that different processes are coordinated within a higher-order scheme, such as proportional reasoning or the concept of confidence interval.
- *Likelihood Fallacy*: "a confidence interval contains the likely values for the parameter. Values inside the confidence interval are more likely than those outside. This fallacy exists in several varieties, sometimes involving plausibility,

credibility, or reasonableness of beliefs about the [actual value of the] parameter" (Morey et al., 2016, p. 105).

- *Margin of error*: can be considered the maximum possible distance between the point estimate and the actual value of the parameter while maintaining the determined level of confidence.
- *Mean*: a measure of center for a quantitative variable.
- *Mean of a sampling distribution*: the expected value of all possible estimates for a given statistic.
- (*Mis)conception*: a word that is historically used to "designate a student conception that produce[d] a systematic pattern of errors" (Smith et al., 1994, p. 119).
- *Observational unit*: the person, animal, object, procedure, etc.
- *Obstacles*: errors and failures connected to prior knowledge originally helpful, but that had become less helpful when applied to larger knowledge domains (i.e. multiplication of natural numbers where multiplication made bigger versus multiplication of all types of numbers where multiplication did not necessarily make bigger).
- *Obstacles of didactical origin*: based on the decisions that teachers and education systems made concerning curriculum or had socio-cultural origins.
- *Obstacles of epistemological origin*: based on necessary decisions made during the didactical transposition of knowledge, which is often required for the development of knowledge.

- *Obstacles of ontogenic origin*: produced from issues that arise from knowledge developed during different stages of a learner's development. A learner can develop ideas based on the neurophysiological limitations and cognitive abilities available to him/her at the time of development.
- *Outcome*: an unknown, but possible value (or attribute) of a random experiment.
- *Parameter*: a numeric summary of the variable for a population.
- *Personal concept definition*: the individually derived or constructed definition of a curricular concept.
- *Pivots*: functions of the random variable such that the distribution of the pivot is free from unknown values of the parameters.
- *Point estimate*: is realized value of a function of a random variable that could be used to estimate a value of the unknown parameter, θ.
    - *Note*: Ideally, this estimate was calculated from an estimator that was the *uniform minimum variance unbiased estimator* of the unknown value of the parameter. *Uniform minimum variance unbiased estimator* (UMVUE) ensured the estimator was unbiased (the expected value of the point estimator of θ equals θ for all possible values of θ) and the variance of the point estimator was less than or equal to the variance of all other possible unbiased point estimators.
- *Point estimator*: a function, $T(x_1, x_2 \ldots x_n)$, of a random variable that could be used to estimate a value of the unknown parameter, θ.
- *Population*: the collection of all observational units of interest.

- *Precision Fallacy*: “the width of a confidence interval indicates the precision of our knowledge about the [actual value of the] parameter. Narrow confidence intervals correspond to precise knowledge, while wide confidence errors correspond to imprecise knowledge” (Morey et al., 2016, p. 104).
- *Probability*: the long-run relative frequency of a random variable.
- *Quantitative variable*: a collection of numeric data on which arithmetic makes sense.
- *Random experiment*: an experiment or trial whose outcome is not predictable in the short term, but for which the long-run relative frequency of outcomes of different types in repeated trials is predictable.
- *Random process*: the process through which random variables exist and are assigned probabilities. These can be considered models that are used to describe real-world data.
- *Random sample*: a sample that was generated through a process that ensures, to the best of the researcher’s ability, that every possible sample from the population had an equal probability of being selected.
- *Random variable*: a variable that is produced from a random experiment.
- *Relational knowledge*: knowledge of the relationship among the components of a confidence interval (i.e. the relationship between confidence level and interval width) (Crooks, 2014; Fidler, 2005).
- *Relative frequency*: the number of data of one outcome divided by the total number of data for all possible outcomes.
- *Sample*: a subset of observational units taken from the population.

- *Sample size*, n: the number of units of interest in the sample.
- *Sampling distribution of a statistic*: the theoretical distribution all possible estimates of a statistic calculated from all possible samples of size n.
- *Scheme*: as a construct to capture the complexity of thinking with higher-order mathematical concepts, was defined by Thompson et al. as "an organization of action, operations, images, or schemes – which can have many entry points that trigger action and anticipations of outcomes of the organization's activity" (2014, p. 11).
- *Shape of a distribution*: a description of the general behavior of the distribution (e.g. symmetric, skewed, uniform, unimodal, bimodal, etc.).
- *Standard deviation*: a measure of variability that was specifically the square root of the squared deviations.
- *Standard deviation of the sampling distribution*: as the theoretical standard deviation of the sampling distribution.
- *Standard error*: as the estimated standard deviation of the sampling distribution. This can be either an approximated value from a simulated sampling distribution or a standard deviation of the sampling distribution estimated from a statistic.
- *Statistic*: a numeric summary of the variable calculated from the data provided by a sample.
- *Theoretical probability distribution*: the formulaic representation of the probability of a random variable.

- *Unbiased estimator*: an estimator for which the expected value (or mean) of the estimator was the value of the parameter for which it was intended to be an approximation.
- *Variable*: a collection of data that could be either quantitative or categorical in nature.
- *Variability*: a measure of the how data values typically deviate from a center value.

APPENDIX B

RECRUITMENT SCRIPT INITIAL EMAIL TO PROFESSORS

Dear Professor:

I am a graduate student in the Department of Mathematics and Science Education working under the direction of Dr. Jennifer J. Kaplan in the Department of Statistics at The University of Georgia. I am emailing to ask for your help recruiting students for participation in my research study entitled “Student Conceptions of Confidence Intervals”. The purpose of this study is to learn how students understand confidence intervals. Would you be willing to email the attached recruitment letter to your students?

Thank you for your time.

Sincerely,

Kristen Roland

APPENDIX C

RECRUITMENT SCRIPT FOR INTRODUCTORY STUDENTS

Dear STAT XXXX students:

I am a graduate student in the Department of Mathematics and Science Education working under the direction of Dr. Jennifer J. Kaplan in the Department of Statistics at The University of Georgia. I invite you to participate in a research study entitled “Student Conceptions of Confidence Intervals”. The purpose of this study is to learn how undergraduate students understand confidence intervals. We have asked the STAT XXXX coordinator, **XXXX**, to contact students enrolled in STAT XXXX in Summer 2019 on our behalf.

In order to be eligible for this research, you must both be:

1) Completed STAT XXXX during the Summer 2019 semester at the **[Name of institution]** and
2) 18 years or older

Your participation will involve a series of interviews. You will be interviewed at most three times during the semester. Each interview will last no more than one hour so your total time commitment will be less than 3 hours. After preliminary analysis of the first interviews, candidates for the second interview will be selected. The participants for the third interview will be selected from the participants who completed the first and second interviews. If you are chosen for a second interview and/or third interview, the researcher will contact you to arrange the next interview. You will receive an incentive of 10 dollars at the conclusion of your first interview. If you are chosen for the second and third interviews, you will be paid 15 dollars at the end of the second one-hour interview, and 20 dollars at the end of the third interview you will have participated.

If you are interested in participating in this study or have any questions about this research project, please feel free to call me (Kristen) at (XXX) XXX-XXXX or send an e-mail to kristen.roland25@uga.edu.

Thank you for your consideration!

Sincerely,

Kristen Roland

APPENDIX D

RECRUITMENT SCRIPT FOR INTERMEDIATE, CAPSTONE AND GRADUATE STUDENTS

Dear [STAT XXXX/ graduate] students:

I am a graduate student in the Department of Mathematics and Science Education working under the direction of Dr. Jennifer J. Kaplan in the Department of Statistics at The University of Georgia. I invite you to participate in a research study entitled "Student Conceptions of Confidence Intervals". The purpose of this study is to learn how undergraduate students understand confidence intervals.

In order to be eligible for this research, you must both be:

1) Currently enrolled in [STAT XXXX/ as a graduate student in the Department of Statistics] at the **[Name of Institution]** and
2) 18 years or older

Your participation will involve a series of interviews. You will be interviewed at most three times during the semester. Each interview will last no more than one hour so your total time commitment will be less than 3 hours. After preliminary analysis of the first interviews, candidates for the second interview will be selected. The participants for the third interview will be selected from the participants who completed the first and second interviews. If you are chosen for a second interview and/or third interview, the researcher will contact you to arrange the next interview. You will receive an incentive of 10 dollars at the conclusion of your first interview. If you are chosen for the second and third interviews, you will be paid 15 dollars at the end of the second one-hour interview, and 20 dollars at the end of the third interview you will have participated.

Your involvement in the study is voluntary, and you may choose not to participate or to stop at any time without penalty or loss of benefits to which you are otherwise entitled. Your instructor will not know whether you participated in this study. Participation in the study will not affect your course performance.

If you are interested in participating in this study or have any questions about this research project, please feel free to call or text me (Kristen) at (XXX) XXX-XXXX or send an e-mail to kristen.roland25@uga.edu.

Thank you for your consideration!
Sincerely,
Kristen Roland

APPENDIX E

FOLLOW-UP RECRUITMENT SCRIPT FOR PROFESSORS

Dear Professor:

I am a graduate student in the Department of Mathematics and Science Education working under the direction of Dr. Jennifer J. Kaplan in the Department of Statistics at The University of Georgia. I am following up to an email I sent asking for your help recruiting students for participation in my research study entitled "Student Conceptions of Confidence Intervals". The purpose of this study is to learn how students understand confidence intervals. Would you be willing to email the attached recruitment letter to your students?

Thank you for your time.

Sincerely,

Kristen Roland

-OR-

Dear Professor:

I am a graduate student in the Department of Mathematics and Science Education working under the direction of Dr. Jennifer J. Kaplan in the Department of Statistics at The University of Georgia. I appreciate your help in recruiting students to participate in my research study entitled "Student Conceptions of Confidence Intervals". While I have received some responses from your students, I would like to attempt to recruit more. Would you be willing to send a second email with the attached recruitment letter to your students?

Thank you for your time.

Sincerely,

Kristen Roland

APPENDIX F

IN PERSON RECRUITMENT SCRIPT FOR INTERMEDIATE STUDENTS

Hello, I am a graduate student in the Department of Mathematics and Science Education working under the direction of Dr. Jennifer J. Kaplan in the Department of Statistics at The University of Georgia. I am here to invite you to participate in a research study entitled "Student Conceptions of Confidence Intervals". The purpose of this study is to learn how undergraduate students understand confidence intervals.

In order to be eligible for this research, you must be:

1) Currently enrolled in STAT XXXX [STAT XXXX, STAT XXXX] at the **[Name of Institution]** and
2) 18 years or older

Your participation will involve a series of interviews. You will be interviewed at most three times during the semester. Each interview will last no more than one hour so your total time commitment will be less than 3 hours. After preliminary analysis of the first interviews, candidates for the second interview will be selected. The participants for the third interview will be selected from the participants who completed the first and second interviews. If you are chosen for a second interview and/or third interview, I will contact you to arrange the next interview. You will receive an incentive of 10 dollars at the conclusion of your first interview. If you are chosen for the second and third interviews, you will be paid 15 dollars at the end of the second one-hour interview, and 20 dollars at the end of the third interview you will have participated.

Your involvement in the study is voluntary, and you may choose not to participate or to stop at any time without penalty or loss of benefits to which you are otherwise entitled. Your instructor will not know whether you participated in this study. Participation in the study will not affect your course performance.

If you are interested in participating in this study or have any questions about this research project, please feel free to call or text me (Kristen) at (XXX) XXX-XXXX, send an e-mail to kristen.roland25@uga.edu.

Thank you for your consideration!

Sincerely,

Kristen Roland

APPENDIX G

CONSENT LETTER

UNIVERSITY OF GEORGIA
CONSENT LETTER
STUDENT CONCEPTIONS OF CONFIDENCE INTERVALS

Dear Participant,

My name is Kristen Roland and I am a student in the Department of Mathematics and Science Education at the University of Georgia under the supervision of Jennifer J. Kaplan in the Department of Statistics. I am inviting you to take part in a research study.

I am doing research on student conceptions of confidence intervals. The primary purpose of this study is to begin to identify the knowledge students have developed concerning confidence intervals. Specifically, I aim to explore two specific questions: 1) what type of knowledge, instrumental or relational, have students constructed when conceptualizing confidence intervals, and 2) how have undergraduate and graduate students structured their knowledge about confidence intervals.

In order to be eligible for this research, you must

1) have completed STAT XXXX at the **[Name of Institution]**
   OR
   be enrolled currently in one of the following: STAT XXXX, STAT XXXX, STAT XXXX at the **[Name of Institution]**
   OR
   be a graduate student currently in the Department of Statistics at the **[Name of Institution]**

AND

2) be 18 years or older

If you agree to take part in this study, you will be asked to participate in a series of interviews. You will be interviewed at most three times during the semester. Each interview will last no more than one hour so your total time commitment will be less than 3 hours. After preliminary analysis of the first interviews, candidates for the second interview will be selected. The participants for the third interview will be selected from the students who participated in the first and second interviews. If you are chosen for a second interview and/or third interview, the researcher will contact you to arrange the next interview. Interviews will take place outside of scheduled class meetings in an empty conference room or classroom at a time that is mutually convenient.

Your involvement in the study is voluntary, and you may choose not to participate or to stop at any time without penalty or loss of benefits to which you are otherwise entitled. Your instructor will not know whether you participated in this study. Participation in the study will not affect your course performance. If you decide to withdraw from the study, the information that can be identified as yours will be kept as part of the study and may continue to be analyzed, unless you make a written request to remove, return, or destroy the information.

There is no expected risk to you during this study. There are no known risks associated with this research. There are some minimal discomforts associated with this research. The discomforts include the stress due to the presence of a video camera in the room or because you may be asked to explain your thinking when you are not sure whether what you are saying is statistically correct. You may experience some anxiety in being asked statistics questions that may make you uncomfortable. You can skip these questions if you do not wish to answer them. We are not interested in right or wrong answers. If you feel uncomfortable, you are free to skip questions or tasks, or discontinue the interview at any time without explanation.

There is no anticipated direct benefit to you for your participation in this study, but the study may allow you to think more critically about confidence intervals. Your participation in the study will help Statistics educators understand how students reason both correctly and incorrectly about confidence intervals.

In order to maintain confidentiality, you will be assigned a pseudonym. All data collected from you will be stored using this pseudonym. Your real name will not be used in any publications or academic presentations: All publications related to this project will use pseudonyms. Research records will be labeled with study IDs that are linked to you by a separate list that includes your name. This list will be destroyed once we have finished collecting information from all participants. The information collected for this study will be only be used for this study and will not be distributed beyond this study.

The interviews will be video-recorded but the camera will be focused on the work you are doing and your gestures. Should your face appear in the recording, your face will be obscured. The video recordings will be transcribed, and both the transcriptions and the video recordings will be used for data analysis. We plan to keep video recordings for five years, and any written work and transcripts indefinitely. These data will be stored in researcher's password-protected personal cloud-based storage and on a portable hard drive.

You will receive an incentive of 10 dollars at the conclusion of your interview. If you are chosen for the second and/or third interviews, you will be paid 15 dollars at the conclusion of the second one-hour interview, and 20 dollars at the conclusion of the third one-hour interview. If you decide to withdraw from any interview at any point, you will not lose the incentive for the given interview (i.e. if you decide to withdraw from the first interview after 15 minutes, you will still receive the incentive of 10 dollars). You will be asked to complete a receipt for each cash payment provided which includes your name and signature. This will be shared with the investigator's departmental business office. If you choose to withdraw consent later, you will not lose your incentive (i.e. the payment you have already received).

If you are interested in participating or have questions about this research, please feel free to contact me at XXX-XXX-XXXX, kristen.roland25@uga.edu. After I graduate in May 2020, please direct any questions or concerns to Dr. Jennifer J. Kaplan in the Department of Statistics (jkaplan@uga.edu). If you have any complaints or questions about your rights as a research volunteer, contact the IRB at 706-542-3199 or by email at IRB@uga.edu.

Please keep this letter for your records.

Sincerely,

Kristen E. Roland

APPENDIX H

INTERVIEW 1 PROTOCOL

Taken from Koklu (2017):

1. Is this the first time you are taking an introductory statistics course (like STAT 2000)? If not, when did you take it before? Tell me your experiences with that class.
2. What other statistics course, if any, have you taken before? Tell me your experiences with those courses.
3. Did you learn any statistics in high school? Before high school? Tell me your experiences with statistics.
4. What is your (intended) major?
5. What are your thoughts about STAT 2000?
6. What are your thoughts about statistics?

Overarching Probing Questions:

- Can you describe this picture to me?
- What is this picture centered at? What is this picture's measure of variability?
- Which sampling distribution is this?
- What is the difference between the standard deviation of the sampling distribution and the standard error of the sampling distribution? Is there a difference?

Task:

Friends, Jamie and Alex, created a group playlist for a long car ride. Since their individual choice in music differed greatly, the pair had agreed to put an equal amount of songs on the group playlist. After listen to 50 songs that were played by the random shuffle, 33 of the songs were songs Jamie added to the playlist. Thinking that Jamie added more songs to the playlist than Alex, Alex questions the proportion of songs on the playlist that were added by Jamie. Suppose the pair wanted to estimate the proportion of songs on the playlist that belonged to Jamie.

1. How should Jamie and Alex do this?
2. Suppose Jamie and Alex wanted to calculate a 95% confidence interval but could not remember how to do that. How should they calculate this confidence interval?
3. How should Jamie and Alex interpret the 95% confidence interval you just calculated?
4. Jamie and Alex cannot remember what the 95% represents in the calculation and the interpretation. How would you remind them what the 95% represents?
5. Jamie and Alex remember an instructor referring to a confidence interval as a range of plausible values. They cannot remember for what the interval is a range of plausible values. The pair also does not understand these terms. How could you help Alex and Jamie understand what the phrase, “range of plausible values” means?
6. If Alex wanted to increase their sample of songs from 50 songs to 150 songs while maintaining the same proportion (66% or 99 out of 150 songs belonging to Jamie), what effect would the change in sample size have on the confidence

interval?, what would you tell Alex about the effect the change in sample size have on the confidence interval?

7. If Alex wanted to calculate a 99% confidence interval instead of a 95% confidence interval with the same sample (33 out of 50 songs belong to Jamie), what would you tell Alex about the effect the change in confidence level have on the confidence interval?

Rationale for Inclusion:

This task was designed to develop a baseline understanding of the concepts and connections participants hold concerning confidence intervals.

- **Question 1:** By asking a general question concerning inference, I aimed to gather information on the participant's initial response to inference.
- **Question 2:** It was my intention to have the students demonstrate their knowledge concerning the calculation of a confidence interval. I, further, intended to ask the student questions about the components of a confidence interval in an attempt to determine what types of connections, if any, the participant had made about the calculation of confidence intervals. This question focused on the procedure of the confidence interval.
- **Question 3:** This question was designed to determine how the student explains what a confidence interval means. I intended to probe the student, using follow up questions, about the different components of the interpretation of a confidence interval. In particular, I was interested to see if the participant had a working definition of the word "confident". I wanted to see how the student describes the parameter of interest. This question was used to determine the depth of

knowledge the student can demonstrate. This question focused on the interpretation of a confidence interval. The follow up questions concerning what the interval represents attempts to isolate any connections the participants had concerning Parameter/Statistic.

- **Question 4:** My intention with this question was two-fold. One, I wanted to determine if the student has a working definition of the word plausible as it relates to statistics. Secondly, I wanted to see if the participant had developed the connection that confidence intervals are the range of plausible values for the actual value of the parameter. Further, as an extension of **Question 2**, this question pushed the participant to explain, in context, what the actual value of the parameter value is. This question focused on the interpretation of a confidence interval.
- **Question 5:** This question was a more directed question concerning the participant's working definition of the word confidence. It was also probing the participant's knowledge concerning the long-run interpretation of the confidence level. This question focuses on the interpretation of a confidence level.
- **Questions 6 and 7:** These questions focused on the relational characteristics of a confidence interval. My intention with these questions was to see if the student can demonstrate the proper relationship between sample size change and confidence interval width and confidence level change and confidence interval width. I also probed the student, using follow up questions, to see if the student has any additional connections concerning the theoretical reason for the relational characteristic.

APPENDIX I

INTERVIEW 2 PROTOCOL

Theme / Research Questions:

- Explore student understanding of specific interpretations of confidence interval.
- Explore student understanding of specific interpretations of confidence level.

Optional Task:

Jamie, a reporter for the Hill Top State University newspaper, has collected the monthly rent from a random sample of 100 HTSU students.

1. What do the following sentences mean?
    - The actual mean monthly rent for all students at HTSU is between \$705 and \$793.
    - The actual mean monthly rent for all students at HTSU falls within \$705 and \$793.
    - The mean monthly rent for the 100 selected students at HTSU is between \$705 and \$793.
    - The monthly rent for all students at HTSU is between \$705 and \$793.
    - The mean monthly rent for samples of 100 students at HTSU is between \$705 and \$793.

Task 1:

Jamie, a reporter for the Hill Top State University newspaper, needs to write a story that discusses the average monthly rent students at HTSU pay for housing. Jamie is looking to estimate the average monthly rent for students at HTSU. After randomly surveying 100 students, Jamie calculates the 93% confidence interval to be ($705, $793).

1. What do the following sentences mean?
    - Jamie is 93% confident that the actual mean monthly rent for all students at HTSU is between $705 and $793.
    - Jamie is 93% confident that the actual mean monthly rent for all students at HTSU falls within $705 and $793.
    - Jamie is 93% confident that the mean monthly rent for the 100 selected students at HTSU is between $705 and $793.
    - Jamie is 93% confident that the monthly rent for all students at HTSU is between $705 and $793.
    - Jamie is 93% confident that the mean monthly rent for repeated samples of 100 students at HTSU is between $705 and $793.
2. What do the following sentences mean?
    - Approximately 93% of the actual mean monthly rent for all students at HTSU is between $705 and $793.
    - Approximately 93% of the mean monthly rent for the 100 selected students at HTSU is between $705 and $793.
    - Approximately 93% of the monthly rent for all students at HTSU is between $705 and $793.

- Approximately 93% of that the mean monthly rents for repeated samples of 100 students at HTSU is between $705 and $793.
- Approximately 93% of all samples of size 100 from the HTSU student body will produce 93% confidence intervals that capture the actual mean monthly rent for all students at HTSU.

3. What do the following sentences mean?
   - Jamie is 93% confident that the actual mean monthly rent for all students at HTSU is between $705 and $793.
   - Jamie is 93% sure that the actual mean monthly rent for all students at HTSU is between $705 and $793.
   - Jamie has a 93% probability that the actual mean monthly rent for all students at HTSU is between $705 and $793.
4. What do the following sentences mean?
   - Jamie is 93% confident that the actual mean monthly rent for students at HTSU is between $705 and $793.
   - The process used to generate confidence intervals will capture the actual mean monthly rent for students at HTSU approximately 93% of the time.
   - There is a 93% probability that the actual mean monthly rent for students at HTSU is within the interval $\bar{x} \pm \left(t_{n-1}^{*} \frac{s}{\sqrt{n}}\right)$.

Task 2:

Jamie talked to a fellow reporter about constructing 93% confidence intervals. Jamie's colleague said that prior to collecting his sample, there is a 93% probability that the confidence interval will capture the actual mean monthly rent of all HTSU students. The colleague continued the explanation by saying that once Jamie collected a sample, the probability of the interval ($705, $793) actually containing the mean monthly rent of all HTSU students is now either 0 or 1.

Task 3:

Students at Mid South State University (MSSU) collected information from all of their students to create a census database. One of the variables they collected information on was the monthly rent for all students at MSSU.

Applet 1: StatCrunch (Data Set – Mid South State University Census Data)

Applet 1: (https://digitalfirst.bfwpub.com/stats_applet/stats_applet_4_ci.html)

Applet 2: http://wise.cgu.edu/portfolio/demo-confidence-interval-creation/

Applet 3: http://www.rossmanchance.com/applets/ConfSim.html

Task 4:

Suppose Jamie wanted to estimate the average amount of money students at HTSU spend on their textbooks per semester. Jamie decides to estimate the average amount of money spent on textbooks by HTSU students with a confidence interval. What considerations should Jamie think about prior to collecting his data?

Rationale:

Task 1: The scenario for this task is based on an example in Kaplan et al. (2018). The choice of a 93% confidence interval was to determine if the participants were able to express their understanding of confidence intervals beyond the typical 95% confidence level.

- **Question 1:** This question contains several statements concerning different interpretations of the range of a confidence interval. The first two statements concern the difference between the word selection “between” and “falls within”. The former is the generally accepted form of the interpretation of a confidence interval because “between” does not imply action on the part of the actual value of the parameter. “Falls within” implies action on the part of the actual value of the parameter, indicating that the actual value of the parameter is not a fixed, but unknown, value. The third bullet implies that the confidence interval is the sample mean is within the interval (with 93% confidence). The fourth bullet describes the confidence interval as the range of the population data. The fifth bullet describes the interval as the range of sample means.
- **Question 2:** This set of sentences discuss interpretations of the confidence level. The first bullet describes range of values for 93% of the population mean. The second bullet describes range of values for 93% of the sample mean. The third bullet describes the range in which 93% of the population data would be within. The fourth bullet describes the range within which 93% of the sample means would be within. The final bullet is a correct interpretation of the confidence level.

- **Question 3:** This question discusses the difference among the words: confident, sure, probability, and chance. The first two words (confident and sure) are typically accepted as equivalent words and are considered correct interpretations. These words are used to discuss the difference between the estimator and the estimate, with the interval being interpreted an estimate with no remaining probability. The last two words (probability and chance) are considered incorrect word choices because there is no longer a 93% probability associated with the interval.
- **Question 4:** This question was given to more advanced students who had seen mathematical statistics. These bullets are a more theoretical way of discussing a confidence level.

Task 2: This question was created to determine if a common statement by participants in Interview 1 made sense to all participants. Additionally, this allowed me to push for more understanding concerning the difference between the estimator and the estimate.

Task 3: This question was created to allow me to use visualizations to gain a deeper understanding of the knowledge each participant had concerning the confidence level.

Task 4: This question was designed to elicit student understanding of the practical implications of the relational characteristics of confidence intervals

APPENDIX J

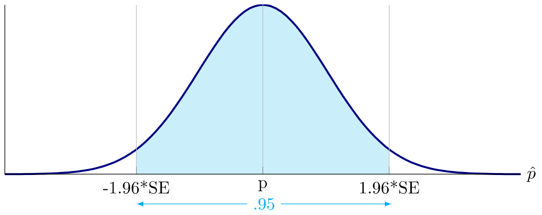
INTERVIEW 3 PROTOCOL

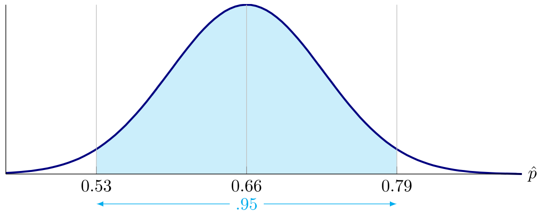
Theme / Research Questions:

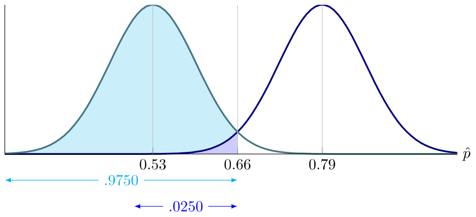
- Explore student's understanding of statistical inference – specifically about developing confidence intervals.

Task 1:

During the first interview, we talked about creating a confidence interval for the proportion of songs on a group playlist added by Jamie. Several participants drew what they visualized as a confidence interval. I have reproduced the pictures here. Do any of these pictures look like what you visualize a confidence interval to look like? If so, which ones? If not, can you describe what you see when you visualize confidence intervals?



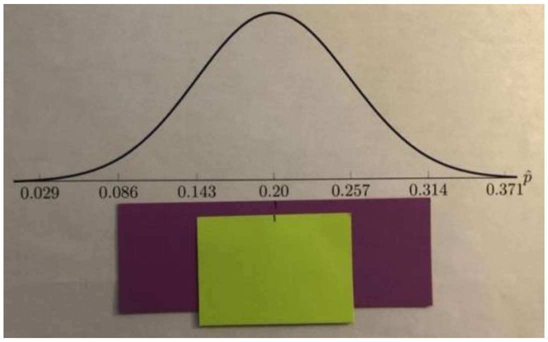




Task 2:

On the Hill Top State University website, the registrar reports that 20% of the 40,000 students are out-of-state students. Suppose that we take 3000 random samples of 50 students each and plot each of the 3000 resulting sample proportions in a histogram.

1. Can you explain to me what the histogram is and describe how it would look?
    a. *What is the histogram an approximation of?*
    b. *What is the shape of this distribution?*
    c. *What value should the distribution centered around?*
    d. *What do you expect the standard deviation of this distribution to be?*
2. What values would contain the central 95% of the sampling distribution?
3. How many standard errors away from the mean of the sampling distribution are each of the values?
4. Using the provided sampling distribution and colored rectangle, what do you think the length of the rectangle represents?



5. Next you are going to take each card and line the center up with some potential values of a sample statistic. You should then determine if the card "covers" 0.20 (what we know to be the true proportion for the population).

- $\hat{p}$=0.36

- $\hat{p}$=0.14
- $\hat{p}$= 0.28
- $\hat{p}$=0.10

6. For the PURPLE rectangle:
   a. What are the maximum and minimum values for a statistic in which the PURPLE card will "cover" 0.20?
   b. What Z-scores correspond to these values?
   c. How often will this sampling distribution yield a statistic where the PURPLE card won't "cover" 0.20?
   d. What is the probability that a PURPLE card centered at a random statistic would "cover" 0.20?

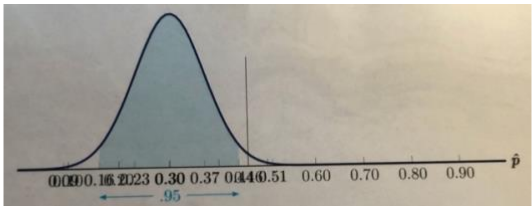
Task 3:

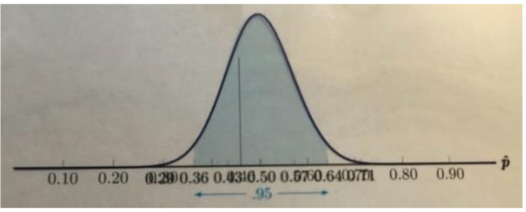
A large Midwestern town has citizens that identify as either liberals or conservatives. We don't know for sure what proportion of the citizens are liberal and what proportion are conservatives. A local news reporter would like to create an interval estimate for the proportion of citizens in the city that identify as liberal and has enough resources to gather a random sample of 50 citizens.

1. Describe the sampling distribution for the sample proportions, the proportion of liberal citizens in each sample of size 50 citizens?

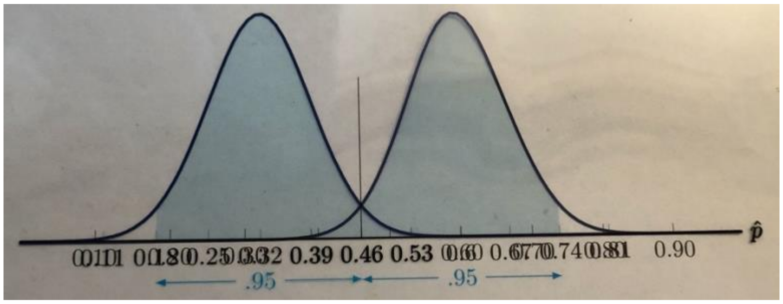
Unlike in the first scenario, we do not know the actual proportion of liberals in the city. Based on what we know about the sampling distribution for the sample proportion and on questions 1-8, answer the following questions:

2. Suppose the local news reporter conducted a random sample of citizens in the city and found that 23 of the 50 citizens of the midwestern town identified as liberal. Decide from which sampling distributions our sample could have been drawn from if we wanted to be part of the region of the sampling distribution would contain 95% of all sample proportions. *[Note: SE was calculated based on the sample proportion, 23/50. Normal distributions centered at p={0.30, 0.32, 0.38, 0.40, 0.44, 0.50, 0.55, 0.60, 0.64}]*



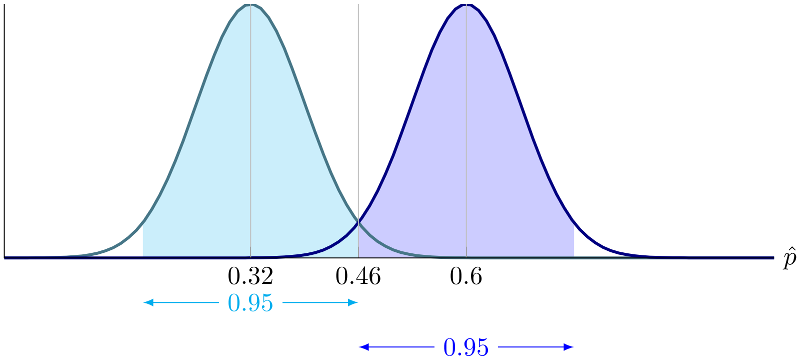


3. What would be the most extreme sampling distributions from which our sample statistic is plausible?



4. What would be plausible values of the actual proportion of liberals in the city, if our statistic was one of the statistics in the region of the sampling distribution would contain 95% of all sample proportions.

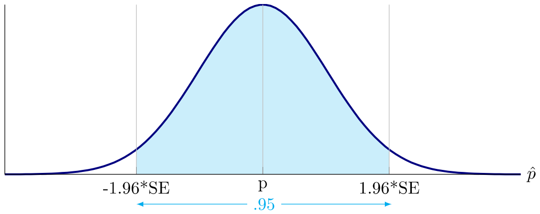
5. Based on the picture below, how could we generalize the above process to create a formula for an interval that would "cover" the actual parameter for 95% of all possible samples?

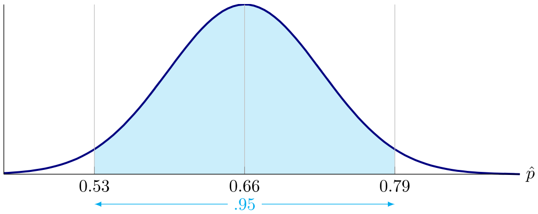


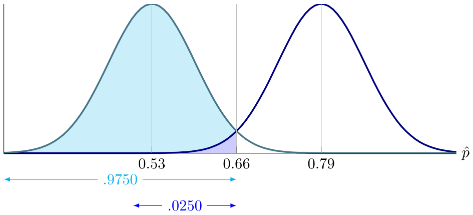
6. How can we interpret the interval you created in question 11? Why do we know this is true?

Task 4:

Now that we have completed this task, do any of these pictures look like what you visualize a confidence interval to look like? If so, which ones? If not, can you describe what you see when you visualize confidence intervals?







Rationale:

Task 1: This question was designed to provide a visual representation of a confidence interval for participants to comment on. This also allowed me to identify a baseline visualization to see if Task 2 would change the participant's ideas of what a confidence interval looked like.

Task 2: This task was designed to be an introduction to confidence intervals. It was created based on the ideas of an activity Dr. Victoria Woodard uses in her introductory statistics course (used with permission, personal communication, January 13, 2019). The context of the task has been modified from Kaplan et al. (2018). The goal of this task is to connect the confidence level with the margin of error, and subsequently the confidence coefficient. This task should also help participants identify where the probability is reflected in the creation of a confidence interval.

Task 3: This task combines the findings from Task 2 with an activity that has been modified from Kaplan et al. (2018). Here, participants connect the ideas of confidence level with the calculation of a confidence interval. It should help participants realized the complexity of the word *confident* and develop a better visualization of the theory behind a confidence interval.

Task 4: This last question was asked to determine if the participant had changed their ideas concerning the visualization of a confidence interval.

APPENDIX K

DATES OF INTERVIEWS WITH STUDY PARTICIPANTS

Participant	Statistical Experience	Interview 1	Interview 2	Interview 3
Joel	Graduate	8/26/19	10/24/19	11/11/19
Liam	Graduate	8/30/19	10/21/19	12/5/19
3*	Intermediate	8/23/19	N/A	N/A
4*	Graduate	8/30/19	10/30/19	11/13/19
Kiara	Intermediate	8/30/19	10/23/19	12/13/19
Aiden	Intermediate	9/6/19	10/21/19	12/2/19
Diana	Capstone	8/29/19	10/22/19	11/21/19
Emma	Capstone	9/4/19	10/23/19	12/5/19
12*	Introductory	8/28/19	N/A	N/A
13*	Introductory	8/28/19	N/A	N/A
14*	Introductory	9/9/19	10/21/19	11/11/19
15*	Intermediate	8/30/19	N/A	N/A
16*	Introductory	9/9/19	10/25/19	11/15/19
17*	Graduate	9/24/19	10/29/19	N/A
Jace	Graduate	9/25/19	10/25/19	11/8/19
Tiana	Introductory	9/23/19	10/23/19	11/8/19
20*	Capstone	10/4/19	11/6/19	12/12/19
21*	Capstone	10/15/19	11/5/19	11/12/19
Brody	Capstone	10/18/19	11/15/19	12/11/19
Logan	Intermediate	10/28/19	11/14/19	12/5/19
Gabe	Capstone	11/5/19	11/14/19	11/21/19

Note. * indicates the participant was removed from the study

APPENDIX L

SUMMARY OF PARTICIPANTS' CONCEPTIONS OF INTERPRETATIONS

For ease of reading, the participants have been grouped by the course from which they were recruited.

Introductory Statistics: Tiana

Tiana was the only introductory student that was analyzed for this study. During the first interview, Tiana's interpretation of a confidence interval appeared to be a confounded sentence that included parts of the sentence for the correct interpretation of the confidence interval and the long-run interpretation of a confidence level: "We are 95% confident that, if we took 100 sample test, the true proportion would fall between this range." She defined the word confident as being based in "sureness," specifically she said: "Not kind of like sure. But we really think that this is the answer. If that makes sense. We feel sure." Tiana also equated confidence with the word chance by stating that the confidence level was the probability the actual value of the parameter was in the interval. On the second interview, however, Tiana explicitly states that probability and chance are not equivalent. Of particular interest is her explanation about probability and confidence. Here, she demonstrated apparently new understanding of the randomness associated with the [actualized] interval:

> it's no longer I want to say that is no longer the probability because it's just like … probability and chance because no longer due to chance anymore because now … he's looking at samples that he's already taken a data from. So, it's no longer …

> like an unsure kind of thing or like a trial kind of thing … like he's already done the data.

With this idea of probabilities and confidence intervals, Tiana firmly stated that probabilities "do not go with confidence intervals" but was able to correctly identify the interpretation of a confidence interval.

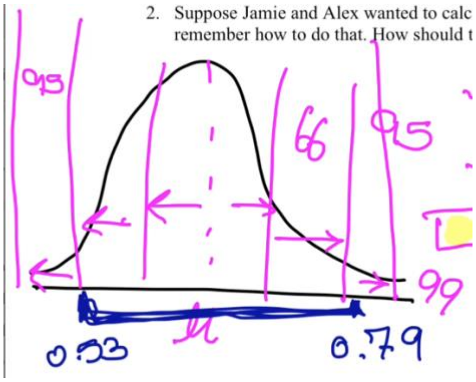
Over the course of the two interviews, Tiana never quite interpreted the confidence level correctly. During the first interview, Tiana wrote that the confidence level was: "the range of where 95% of the true proportion of Alex's (sic) song's in the playlist." This statement implied that there were either multiple true proportions of Jamie's songs on the playlist or that 95% of the actual proportion of the songs on the playlist is with the range (i.e. if the actual proportion of songs is x, then .95 times x is within the range). As with the interpretation of the confidence level, Tiana was able to identify a potentially correct interpretation. During the second interview, she eliminated all of the statements about the interpretation of the confidence level, including the correct long-run interpretation. Her discussion about the correct interpretation indicated that she has a surface level understanding of the long-run interpretation of a confidence level. Initially, Tiana selected the correct interpretation as correct, but she got confused by the multiple "93%" in the statement. She was able to state that the interpretation should be that: "if you did your sample 100 times, the 93% of them will produce the confidence intervals. Basically, I know your actual mean would have been in those 93." Tiana ended up removing the second quantifying 93% because "this 93% [keeps] coming back up again. I don't think we do that. I think a 93% just says here [indicating the first 93%]." When pushed on whether or not the confidence level has an effect on the percentage of intervals that capture the actual value of the parameter, she instead discussed the connection between the confidence level and the confidence interval width:

> Yes, that's where you get like, the alpha thing. Right and Like it would be like .05 and then that would mean that your confidence interval is like 95%. … So, with a higher confidence level you have a more accurate or stronger confidence it makes your confidence interval smaller.

Tiana was able to procedurally find the confidence coefficients for a given confidence level but did not demonstrate any deeper connections to statistical concepts needed to derive confidence intervals. Further, she visualized the confidence interval by placing it on a normal distribution centered at the "mean" (which in this example should have be the true proportion), but with the interval drawn at 2 standard deviations (blue lines), see Figure 12. This suggests that she has confounded several images that are often used in introductory statistics: a normal distribution for the population (Figure 13a), a normal distribution for the sampling distribution (Figure 13b), and the normal distribution calculator for finding confidence coefficients (Figure 13c).

Figure 12

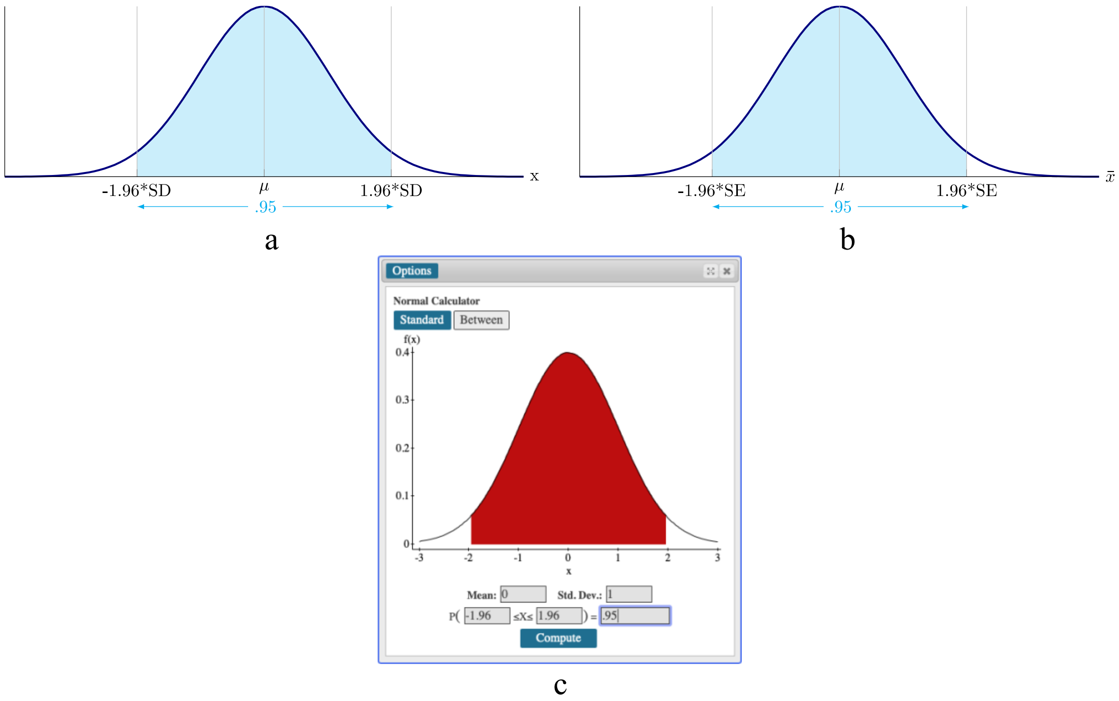
Tiana's Drawing of the Relationship Between the Confidence Interval and the Confidence Level



2. Suppose Jamie and Alex wanted to calc
remember how to do that. How should t

Figure 13

Exemplars of Confounding Visualizations of Confidence Intervals



Mean: 0 Std. Dev.: 1

P(-1.96 ≤X≤ 1.96) = .95

Note. Figure c: From *Normal Calculator Applet*, by StatCrunch, 2019, (www.statcrunch.com), 2019 by Pearson. Reprinted with permission.

Intermediate Statistics

There were three intermediate statistics students that were analyzed for this study: Kiara, Aiden, and Logan. Their conceptualizations of the interpretations of the confidence interval and interpretations of confidence level are provided below.

Kiara

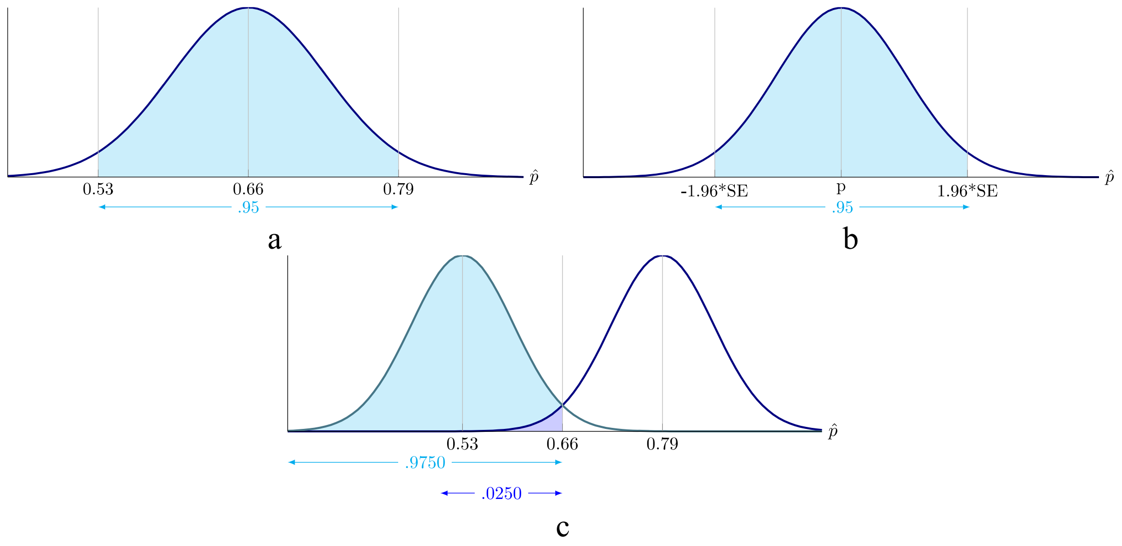
During Kiara's first interview, she stated that it had been over a year since she had taken a statistics course. At the time of the first session, the intermediate statistics course had just started. The growth in her use of statistical terms and statistics concepts improved greatly between her first and second interview. During the first interview, it was unclear if Kiara recalled the difference between statistical terms, such as: parameter, statistic, sample, population. On the second interview, these terms were more clearly defined in her conversations. Kiara's initial interpretation of a confidence interval was: "95% confident of the sample statistic, the sample values you get from your sample." When pushed to describe the values, Kiara stated that she meant: "95% confident that 95% of the actual parameters, the values that you pull from the population fall within the interval." During the first interview, Kiara admitted to having "never thought about it [the word confident] like that I kind of just took it in as it was in the class" and defined the word confident to mean more likely than not. Of particular interest is how Kiara described the "more likely than not:"

> I want to say more likely than not. But I guess that's like, 51% Isn't that great? … I think it means you're almost super certain, there's like a 5% chance … Not I'm on that same numbers, even though there's a 95.

By equating the confidence level of 95% as being "super certain" and "more likely than not", it suggests that Kiara has the conception that a probability greater than 50% equals certainty of happening. On the second interview, however, Kiara discussed the word confident to mean sureness. The use of sureness was still described as being "more sure than not or more confident or not," similar to her explanation during Interview 1. This time, however, Kiara stated "I don't really know how to explain that more. And it's been explained to me more." Therefore, I hypothesize that Kiara does not have a well-defined personal definition of the word confident within the context of confidence intervals. Unlike Tiana, Kiara did not attempt to draw or otherwise visualize the confidence interval or confidence level during her interviews. When presented the three images of different visualizations of confidence intervals in Interview 3 (see: Appendix , Task 1), Kiara did select the visualization of the confidence interval on the normal distribution centered at the statistic (see Figure 14a) as her top choice and the visualization of the interval on a normal distribution centered at the actual value of the parameter (see Figure 14b) as her second choice.

Figure 14 Visualizations of Confidence Intervals Used in Interview 3

Visualizations of Confidence Intervals Used in Interview 3



Kiara never selected or stated an interpretation of the confidence level that implied the long-run interpretation of a confidence level. Instead, Kiara reiterated the interpretation of the confidence interval. During her discussion of the statement that was the correct interpretation, Kiara removed all language that was associated with the long-run interpretation of the confidence level, calling it “weird additional information, but I don't know how you could like come up with that, at least off the top of my head. I don't know.” Instead, Kiara discussed the confidence level as it relates to the confidence interval width. This implies that Kiara may only see the confidence level as it relates to the calculation of the confidence interval and the likelihood of the interval capturing the actual value of the parameter rather than deeply connected to curricular concepts, such as sampling distributions and coverage probability.

Aiden

At the time of Aiden's participation in the study, it had been a while since his last statistics class, and, similar to Kiara, was in the first few weeks of the start of the intermediate course from which he was recruited. One possible limitation that will be discussed in the next chapter is the context of the first interview. The scenario suggesting that Alex wanted to estimate the proportion of songs on the playlist that belonged to Jamie because Alex wanted to see if Jamie had loaded more songs on the playlist, which implied a possible null hypothesis (p=0.50). Aiden, whether by the nature of his understanding of statistical inference or by the nature of the context of the problem, interpreted and discussed many of the questions as if they were suggesting a hypothesis test analysis. This was particularly evident when Aiden attempts to visualize the confidence interval. In his visualization, the distribution was centered at the hypothesized value of p=0.50, around which a confidence interval was created. Aiden pictured determining if the $\hat{p}$ of the problem, 0.66, was within the calculated interval. When presented with the three visualizations of the confidence interval on Interview 3, Aiden still wished to see the hypothesized value of p=0.50 on the visualizations. Although, he was drawn to the picture with the two normal distributions, he eventually chose Figure 14a, the interval centered at the statistic, with Figure 14b being a close second since it did not have any of the numbers plugged in.

When Aiden initially interpreted the confidence interval, he began by stating the interpretation was "there's a 95% chance that the true value is within that range. 95%

chance always sounds wrong, because 95% chance, I'm pretty sure it's something you're never supposed to say." Aiden continued this explanation with a glimmer into his understanding of probability: "unless you're literally saying like there's a 95% that's something happening. 95% chance of rain is not true, it's 95% of models say that it will rain. So 95% chance sounds wrong to me." As Aiden recalled the word confident, he implied that confident was the right word, but that it felt wrong because it seemed like it was an "arbitrary declaration:"

> like you're 95% sure, like, I'm 99.9% sure that I'm you know, going to be late for class. It's not that in you know, 100 situations 99, and a half of them seemed like I will be late. So that way of phrasing it tends to be misleading in my head. But I'm pretty sure that it is we're 95% confident that the true value of the population is within those bounds.

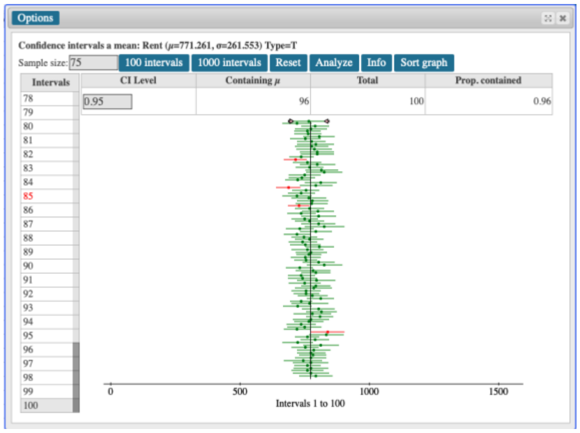
When discussing the difference in the statements containing confident, sure, and probability on the second interview, Aiden stated that confidence was "the likelihood of 100% certainty." It appeared Aiden's definition of confident/confidence contains the idea that it is a likelihood of certainty. This appeared to affect his understanding of the confidence level, itself. In Interview 1, Aiden described the interpretation of the confidence level as a probability, "I would say that it is 90% of the time, or we are 95% sure that the true value's in there." He continued his explanation by stating that 95% of the time was not the correct way to explain the confidence level "because the population parameter is a theoretically like a fact, right? That is a that is a number that will not change does not change it is population parameter." Aiden then formally defined the confidence level by using statements typically associated with the interpretation of the confidence interval:

> I would say that it is 90% of the time, or we are 95% sure that the true value's in there, because you can't, you can't get a different value of the population parameter. … So if I had to say, Jamie, Alex, trust me here, here's what it is. It would be we are 95% confident that the true value of the population is within these bounds.

Aiden, however, did not appear to be comfortable with this conclusion because he continued by discussing the confidence level and the term confident using his definitions imply a "self-fulfilling prophecy …you are 95% confident, because you're going to see that that result 95% of the time. … because you can mathematically ensure that 95% of the time, you'll see a result in looking for." On the second interview, Aiden disagreed with the correct interpretation of the confidence level. In this discussion, Aiden presented a different understanding of the confidence level. Here, he suggested that the confidence level was related to the sample and to the fact that if the sample is a representative sample, the interval should contain the actual value of the parameter. When faced with the StatCrunch applet for confidence intervals (see: Figure 15), Aiden eventually agreed with the statement that approximately 95% of the repeated samples would capture the actual value of the parameter. It is unclear from this discussion if Aiden meant the statement that referred to 93% of repeated samples would be within the actualized interval, or the statement that referred to approximately 93% of all samples of size 100 would produce 93% confidence intervals that would contain the actual value of the parameter.

Figure 15

StatCrunch Confidence Interval Applet



Confidence intervals a mean: Rent (μ=771.261, σ=261.553) Type=T

Sample size: 75

CI Level	Containing μ	Total	Prop. contained
0.95	96	100	0.96

Intervals
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

Note. From *Confidence Interval Applet*, by StatCrunch, 2019, (www.statcrunch.com), 2019 by Pearson. Reprinted with permission.

Logan

Logan was recruited to the study during a second round of recruitment, meaning his first interview was further into the semester than Kiara and Aiden. Additionally, he was a year behind Kiara and Aiden and did not imply as long of a break from statistics classes as Kiara and Aiden did. His interviews were more consistent in terms of recalled statistical knowledge than Kiara and Aiden's were. In the first interview, Logan described the interpretation of a confidence interval as "they're 95% sure that Jamie's real share of songs is within that interval," indicating the actualized interval. Logan believed that sure, probability, and confident were all used by his instructors and meant the same thing. During the second interview, Logan stated that confident, sure, probability, and chance all meant the same thing. This time, however, he stated that sure seemed to imply "spit balling," that the percentage associated with the word sure (i.e. 95% sure) was just a guess. Like Kiara, Logan did not draw a picture to visualize a confidence interval on either of the first two interviews. On Interview 3, Logan selected both Figure 14a and Figure 14b. He preferred Figure 14a because the values were on the graph. He was concerned about the "1.96*SE" on Figure 14b because that was the margin of error, not a point on the graph.

When Logan first interpreted the confidence level, he described the level as the "sweet spot" choice between confidence interval width and confidence level. He elaborated this description by saying that "*95% of the time on 95% of the trials run playlist made whatever that Jamie's proportion would be in that interval* [indicating the calculated interval]." On the second interview, however, Logan did choose the correct interpretation of the confidence level by saying "*So I think this can be a correct interpretation, but it's just very wordy*." Finally, while discussing Jamie's Colleague (Interview 2, Task 2, see: Appendix), Logan disagreed with the captures/not capture aspect of the question, instead he stated that the probability remains .93. This indicates that Logan may not have a well-developed understanding of the confidence level. Instead, like Kiara, he seems to understand that it is required to calculate the confidence interval but is not able to describe its true purpose in the interpretation of a confidence interval.

Capstone Students

Four capstone students were analyzed for this study: Diana, Emma, Brody, and Gabe. This section contains analytical summarizes of their conceptualizations of the interpretations of confidence intervals and interpretations of confidence levels. The participants are discussed in no particular order.

Diana

During the first interview, Diana's interpretation of the confidence interval consisted of the long-run interpretation of the confidence level, rather than the traditional statement for the interpretation of the confidence level. Specifically, she stated: "Uh I believe the way I was taught is 95% of the time it will capture the true value." But, she was quick to state that this was not in terms of a probability: "Not that there's a 95% probab..., like not that it's a probability of it being in there, but that if you ran you know, the simulation ever, hundred times 95 of those would have the true population parameter." While some researchers would argue that the interpretation that Diana used in Interview 1 was the correct interpretation (see the Chapter 2, section: *General Interpretations of Confidence Intervals*), Diana never produced an interpretation for the actualized interval. This may be due, in part, to the fact that Diana was unable to compute an interval during the first interval because, on the second interview, Diana did agree with the correct interpretation for the confidence interval. Despite not stating the word confident in her interpretation of the confidence interval, I asked Diana what the word confident meant to her. Diana initially found it difficult to define without using the word confident, instead she said, "how like, how good you think whatever you've gotten is in representing whatever you're trying to estimate … like 95% is kind of more quantifying that strength." This may suggest that despite selecting the correct interpretation from a series of statements on the second interview the statement she selected may be internalized but is not a statement for which she has a well-developed definition.

Additionally, Diana does not have a well-developed visualization for the creation of the confidence interval. On the second interview, she admitted to "be honest blanking on how the sampling distribution relates to the population distribution and because I think you can only relate it with like certain conditions are met. That link is broken in my head." Like all of the intermediate students, Diana never drew a picture to demonstrate her understanding of the confidence interval. Her confusion on the statement about the interval containing repeated samples ("approximately 93% of that the mean monthly rents for repeated samples of 100 students at HTSU is between $705 and $793"), however, may provide evidence into her thinking about the construction of the confidence interval. Initially, Diana stated that the statement made sense:

> maybe even more sense, then some of the other like, normal interpretations of the confidence intervals. Because to me that that captures that if you did repeated samples, because it's only saying approximately 93%. So it's not saying like, you actually did all the samples, but if you did do repeated samples of 100, 93% of those would have a mean, monthly rent that's between these numbers like that actually makes alot of sense to me.

After I reminded her that the interval was a confidence interval, she eventually disagrees with the statement but states that "it's close." This suggest that Diana has, in her head, an image of the confidence interval that is similar to Figure 14a. During the third interview, she confirmed this hypothesis by stating that she pictured Figure 14a for a specific situation and Figure 14b for a more general situation.

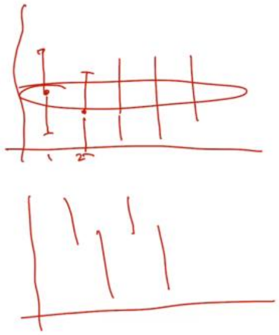
As was just discussed, Diana consistently used the long-run interpretation of a confidence level to interpret the confidence interval. Thus, she was able to use this interpretation when trying to discuss the interpretation of the confidence level. She clarified her previous long-run interpretation statement by saying "if you pull a hundred samples and calculate those proportions. You calculate the you calculate confidence intervals for each of those sample draws 95% of those sample draws would contain the true the true population parameter." She was also quick to explain the reason why probabilities were no longer associated with an [actualized] interval was due to the fact that the [actualized] interval either did or did not contain the actual value of the parameter, so the probability of it containing the actual value of the parameter was either 0 or 1. She also drew a diagram to demonstrate her understanding of the long-run interpretation of the confidence level, but she used this diagram, see in Figure 16, to explain that the overlap in intervals provided strength to your estimation and non-overlapping diagrams would indicate a weaker estimation. Despite demonstrating good recall of the long-run interpretation of the confidence level, Diana was shocked by the StatCrunch applet demonstrating the long-run interpretation of the confidence level. The StatCrunch Applet, see in Figure 15, created 100 95% confidence intervals based on 100 samples of size 75 from a provided "population" data set, a data set that contained population data from a fictional university titled MidSouth State University based on a real data collected from a large southern university. The applet displayed which confidence intervals captured the actual value of the parameter. Diana's response was:

> Interesting. I've never seen anything like that. And I kind of assumed that it was like, a hard and fast rule. So that's surprising. I didn't actually know that like. So then what is the five like, I'm now curious, like, where does the five come from? Why is it showing up? Have questions for my stats teachers.

This indicates that while Diana may have an internalized statement concerning the interpretation of the confidence level, she has not made deep connections to necessary curricular concepts to understand the meaning of the statement.

Figure 16

Diana's Representation of the Long-Run Interpretation of the Confidence Level



Emma

Emma interpreted the confidence interval using the traditionally correct interpretation. Her understanding of the word confident, however, is a bit troubling. Emma believes confident, sure, probability, and chance are all synonyms, which leads Emma to have a probabilistic understanding of the actualized interval. During the first interview, Emma stated that "95% of the time, the true proportion of songs that Jamie load onto the playlist would be somewhere within our interval." Despite this, she defined the word confident using the same idea as Diana, that the interval either captures or does not capture the actual value of the parameter: "well, the true proportion is either the true proportion meaning how many songs Jamie really loaded is either in the interval or not? Of course, it's impossible to know. So, confidence, meaning what we believe."

Of particular interest to me is the fact that Emma holds simultaneous conceptions that are in conflict with each other. Emma visualizes the confidence interval by drawing a normal distribution centered at the statistic with the interval around it (Figure 7c). At the same time, she can conceptualize the sampling distribution centered at the unknown value of the parameter. This was evident by an explanation Emma provides on Interview 2. When describing her understanding of Jamie's Colleague's statement about the probability associated with the creation of a confidence interval prior to collecting a sample (Interview 2, Task 2), she began to discuss the choice of selecting a critical value [confidence coefficient], which included discussing the idea that 93% of a distribution is within the critical values chosen for the calculation of the confidence interval. Emma draws a normal distribution, indicated critical values [confidence coefficients] and wrote 93% on the drawing (see Figure 17a). Because this was a different picture from what she drew previously, I pressed Emma on her drawing. She continued by stating that the shaded region contains 93% of the samples: "I'm pretty sure that these normal distributions that I'm drawing are sampling distributions, in which case they contain a bunch of sample means from in this case and n equals one hundred [drew Figure 17b]." I asked Emma what the sampling distribution was centered at, to which Emma responded:

> Well, it's centered at the mean [meaning μ], but it I mean, we can flip back between centering our distribution around the sample mean and the true mean. I think it's easier when we are centering all of our confidence intervals around the sample mean, it's easy to think of it that way, which is when I did the .93 right here [indicating Figure 17a]. This would be like the true mean right here [adding μ onto Figure 17a to get Figure 17c]. And then here's our 7% of samples that

> we're going to take from our population [adding $\bar{x}$ onto Figure 17a to get Figure 17c] that aren't going to contain the true mean in the interval.

The pictures in Figure 17a, b, and c contradict a picture Emma had drawn previously in the interview: Figure 17d. She drew this picture to explain why 93% of the time the confidence interval captured the actual value of the parameter and 7% of the time it did not. This picture indicated that the sampling distribution was centered at the statistic, the endpoints of the interval were the locations of the critical values she drew in Figure 17 a and b, and μ could be placed anywhere. She explained the drawing as she was making it. She began by stating properties of the normal distribution: that we know it is more likely than not that the collected sample will produce a sample mean that is close to the true population actual value of the parameter. After drawing the normal model in Figure 17d, she stated:

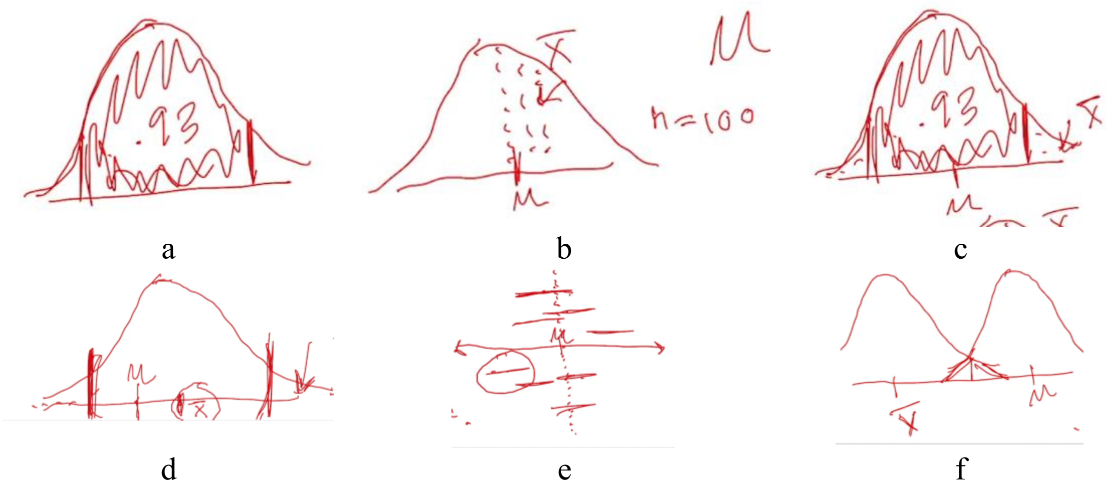
> Let's say this is sample mean [drew normal curve and indicated $\bar{x}$ in Figure 17d] … [μ is] most likely is close to the sample mean … of course there's how we constructed it [meaning the interval] … This was the cut off we made for 93% [drew vertical lines in Figure 17d] that it [meaning μ] could be out here in the tails [drew arrow in Figure 17d] which is unlikely but possible. So just like - the longer - on average doing this a bunch of times, 93% of the time it [meaning μ] will be in here [indicating the main body of Figure 17d] and 7% of the time it will be in one of the tails [indicating the tails on Figure 17d].

I asked for clarification on the percentages Emma described. Unlike in Figure 17 a and b, where Emma indicated 93% of the samples would be within the distribution, and Figure 17c where she indicated the distribution should be centered at the actual value of the parameter, Emma stated that the 93% represented the percent of time the actual value of the parameter would fall within the distribution:

> if we centered the normal distribution to be around the sample mean 93% of the time the true mean would fall somewhere [drew the μ in Figure 17d] in the interval that you created from here to here [indicating the vertical lines in Figure 17d], where 7% of the time it would be somewhere out here in the tails.

I attempted to bring the conflicting ideas to Emma’s attention. Emma appeared to not have an issue with the conflicting ideas. Instead, Emma drew Figure 17f, which appears to be a drawing typically associated with the instruction of hypothesis testing in mathematical statistics courses.

Figure 17: Emma's draws from Interview 2



a b c

d e f

Another source of conflicting mental images took place during Emma’s continued explanation of Figure 17d. When Emma drew Figure 17d, she stated there was a sample mean and a population mean that would “fall” within the diagram she drew. I wanted to gain a better understanding of Emma’s conception of the relationship between the sample mean and the population mean and asked if there were multiple actual value of the parameters (μ). Emma responded by stating that there is one population mean and multiple sample means. Her explanation, however, implied that there is a 93% chance that the multiple sample means would lie within the actualized interval:

> To use like to think of the long run probability where 93 out of 100 will contain the true meaning and the others won't. That would be taking the sample mean from the same population over and over and over because - of course for each interval, it's either in there it's not, but in the long run, there's a 93% chance that it's in there each time [indicating the current interval].

When asked again about whether there was one population mean, or multiple population means, Emma stated: "There's only one. And we don't know what that number is [meaning μ]. But every time we take a sample, even though our sample mean will probably change a little bit, this true mean - the population mean's always staying the same." Here, it is hard to understand Emma's description. She has an idea of a hypothetical sampling distribution centered at an unknown value of a parameter for a particular variable of interest from a population that contains multiple sample means produced from taking samples from the population of interest. But, in the description of the image of the sampling distribution centered at the statistic, she seemed to imply the idea that the statistics vary around the actualized statistic at the center of this distribution (Figure 17d). She further demonstrated these conflicting, but solidly engrained images, by demonstrating the confidence level through the drawing of Figure 17e. As a reminder, Emma had just described a situation where the sample means would vary around the actualized statistic. She continued by drawing Figure 17, which showed the sample means varying around the actual value of the parameter:

> This is a number line [drew horizontal line in Figure 17e]. And this is the true mean whatever that is [drew μ in Figure 17e]. We're not sure what number it [meaning μ] is maybe we would take some samples and most they'd likely going to be close. But, you know, sometimes we'll just have a rare sample that's kind of far away [put random dots on Figure 17e]. And then if all these dots are $\bar{x}$, they'll have the same width around them [drew lines on each of the dots in Figure 17e]. Confidence intervals. [drew dotted vertical line at μ] I didn't draw a 93% confidence for this one, but we would expect 93% of them to contain the true mean, meaning that they would cross this line whereas confidence interval like this [circled line that did not cross the dotted line in Figure 17e] is one of the 7% times where we didn't predict the true mean correctly.

The contradictory images for the sampling distribution could explain her conflicting statements about the interpretation of the confidence level.

During Interview 1, Emma provided three different interpretations of a confidence level. The second explanation coincided with the imagery in Figure 17e, which was the correct long-run interpretation of the confidence level:

> That if we made if we took 100 samples of size 50, from the 1000 Song Playlist, and we made 95% confidence intervals for each one of the samples, we would expect, roughly 95% of the confidence intervals to contain the true proportion, the remaining five out of the 100 samples and not contain the true proportion

Emma's first explanation, however, coincided with her conceptualization of a confidence interval as Figure 17a and Figure 7c by relating the confidence level with the width of the confidence interval. The last explanation of the confidence level that Emma provided during the first interview coincided with her definition of confident: “95% of the time, the true proportion songs that Jamie load onto the playlist would be somewhere within our interval.” As a result of her mental images of the confidence interval and her understanding of the word confident, Emma appeared to see the confidence level as a measure of probability. This probability did appear to be connected to the coverage probability since she can describe the connection between the confidence level, confidence coefficient, and the relationship to the proportion of statistics within the confidence coefficients (as seen in Figure 17c).

Brody

Brody’s conceptualization of confidence intervals was the most conceptually based out of all of the participants in this dissertation study, including those not fully analyzed. Not only was Brody able to clearly articulate interpretations for the confidence interval and interpretations for the confidence level, Brody was also able to explain, conceptually, why the interpretations existed. Like Emma, Brody had parallel conceptions but had not connected some of these conceptions together until pressed during the conversations that took place during the interviews for this dissertation study. To begin, Brody interpreted the confidence interval using the traditional sentence associated with the interpretation, although he referred to this sentence as “formulaic”: “We are 95% confident that the true samp, or the true population proportion of songs that Jamie added is between .5287 and .7913.” He continued by explaining that people often understood this sentence as a probability, which he stated was incorrect. Brody used the long-run interpretation of the confidence level and the probability of the actualized interval either capturing (1) or not capturing (0) the actual value of the parameter to explain how the interpretation of the [actualized] interval does not a include probability. Brody was also the only participant who disliked the use of the word “sure” as an equivalent to the word “confident” in the interpretation of a confidence interval, stating that the word sure more closely implied probability rather than the intended meaning:

> I think 93% sure sounds like it saying a 93% probability. So. like it's saying like there's a 93% chance which we're going to see in a second is like not necessarily correct. So, I think the confidence level shows more that that's not what that means. Whereas with when you say like 93% Sure, it kinda is a little bit more vague.

Brody spoke of the word confident as a quantifiable measure. He saw the word confident as indicating that a confidence interval was used and was clearly able to articulate that he did not interpret the word confidence in a probabilistic way, unlike the majority of the participants in this study. Brody was also able to clearly articulate the traditional statement for the interpretation of the confidence level: “I'd say 95% represents [writes what is saying], obviously the confidence level, but the fact that if you were to construct n 95% confidence intervals with the same population, approximately 95% of those intervals would contain the true parameter p, p_j [where p_j represents the proportion of songs on the playlist that belonged to Jamie].” Brody can explain this interpretation with two different examples. Initially, Brody described the long-run interpretation of the confidence level using diagrams similar to Figure 15, the Confidence Interval Applet in StatCrunch. He can visually describe how approximately confidence level percent of intervals would capture the actual value of the parameter. In the following paragraphs, a second visualization and explanation of the coverage probability.

What makes Brody unique among the participants for this dissertation study is his ability to conceptually derive the concepts and interpretations he was using. Brody was able to describe the process through which the confidence coefficients are selected (procedurally explain the selection of z-scores based on the probability within the normal distribution see Figure 13c) but also explain the connection between the standard normal distribution and the theoretical sampling distribution centered at the unknown value of the parameter. This was evident in his discussion about why the confidence level percentage of intervals captured the actual value of the parameter. In this explanation, Brody demonstrated that he understood the long-run interpretation of the confidence level but had not necessarily connected it to the curricular concepts underlying the interpretation. The ease with which Brody was able to talk through this explanation demonstrated that Brody had the ideas already within his concept image, he had just not made the connections. Brody began by identifying that the distribution that was required for the current context (normal approximation to the binomial distribution, a sampling distribution for $\hat{p}$) would be a normal distribution. He started by identifying the confidence coefficients for the current situation (95% confidence level) but paused for approximately 30 seconds, thinking. When he resumes, he states that the true sampling distribution would be centered at the unknown value of the parameter, p, and began to draw Figure 18a. Brody continued by stating that the standard error of the sampling distribution would be 0.066667, which was incorrect for the true sampling distribution. The standard error Brody calculated, 0.066667, was the estimated value assuming a $\hat{p}$ value of 0.66. The standard deviation of the sampling distribution would be unknown if the value of the parameter is unknown. Eventually, Brody realized this and was able to state that the standard deviation of the sampling distribution [standard error, as Brody stated] was unknown:

> no this is not going to be the standard error. We don't know that they're there for this because we don't have the true p. Anyway, so this $\hat{p}$ for is going to have a standard error, right that based on that standard error that comes from that $\hat{p}$.

The description continued with the explanation that within the sampling distribution are $\hat{p}$ and that 95% of the data are within 1.96 times the standard error of the unknown value of the parameter. Brody stated that the length of this distance (1.96 times the standard error) is the margin of error: “the values out here that contain 95% of this data, right, is that negative 1.96, times the standard error and positive 1.96 times the standard error so that's essentially the margin of error on each on either side.” From these statements, Brody makes the following conclusion:

> So, when we find a $\hat{p}$, right, … So each of these, so let's say we choose a $\hat{p}$. Right? So there's our $\hat{p}$, we know that that $\hat{p}$ is going to have … a standard error, right that based on that standard error that comes from that $\hat{p}$. Right? We're going to create a confidence interval. … Then 95% of your $\hat{p}$ are going to be such that the spread of that interval [draws the line indicating the distance between the left vertical line and p on Figure 18c], right, contains p and so there is 95% of all possible $\hat{p}$ contain p on their interval. And so, I guess that's the best way I can say to explain that.

Here, Brody was able to connect the idea behind the length of the margin of error with the confidence level and the reason why the confidence level percentage of confidence intervals capture the actual value of the parameter. An interesting part of this conversation centered around the initial confusion of the difference between the estimated standard error based on a $\hat{p} = 0.66$ and the theoretical standard deviation of the sampling distribution. Brody appeared to want to make the claim that 95% of the confidence intervals would capture the actual value of the parameter because 95% of the $\hat{p}$'s would be contained within the region between the two vertical lines on Figure 18c. He was, however, faced with the discrepancy between the estimated standard error, which he noted will change based on each $\hat{p}$ value, and the true margin of error that he had drawn in Figure 18c:

> Anyway, so this $\hat{p}$ for is going to have a standard error, right that based on that standard error that comes from that $\hat{p}$. Right? … I guess it assumes that it'll average this out because so a $\hat{p}$ on this side [writes $\hat{p}_1$ on the normal distribution in Figure 18d] is going to have a standard error of [writes SE$_1$ formula in Figure 18d] … and so the that $\hat{p}$ [indicating $\hat{p}_1$] it's gonna be different than then a $\hat{p}$ [writes $\hat{p}_2$ on the normal distribution in Figure 18d] Up here, right so this is [writes SE$_2$ in Figure 18d] be (mumbles) so there's different $\hat{p}$'s are gonna give you different margins of error right.

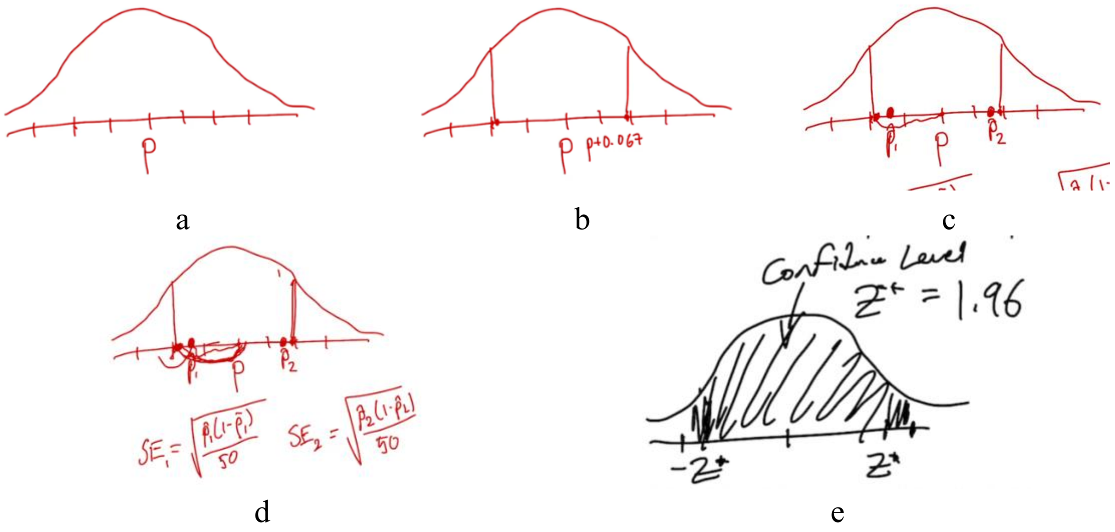
Brody conceded to this internal conflict with the same frustration that teachers feel when trying to teach the concept of coverage probability: "So I feel that this would be really easy if they always had the same standard error be easy to explain because if the margin of error is this distance right.*"* He finished this explanation by stating that this conflict between the estimated standard error and the theoretical standard deviation of the sampling distribution must be the reason the interpretation includes the word "approximately":

> And so if you're using the sample ones, it's probably not that far off would be my guess of how we can still say that. Approximately 95%. So obviously, like, if you get a sample of 100 things, you're not always gonna get exactly 95. So I guess that's why it's approximated, but yeah, that would be my best explanation for that.

The above statement is actually correct, but Brody does not appear to fully believe his conclusion. His ability, however, to discuss these ideas demonstrate the depth of his conceptual knowledge while simultaneously demonstrating that he has not thought about how the estimated standard error may affect the coverage probability of a confidence interval.

Figure 18

Brody's Drawing for His Explanation of Confidence Level



a

b

c

d

e

The next time Brody used the above explanation to discuss the statistical concept of coverage probability was in response to the following statements in Question 4 on Interview 2 (see: Appendix):

1. The process used to generate confidence intervals will capture the actual mean monthly rent for students at HTSU approximately 93% of the time.

2. There is a 93% probability that the actual mean monthly rent for students at HTSU is within the interval $\bar{x} \pm \left(t_{n-1}^{*} \frac{s}{\sqrt{n}}\right)$.

Brody described statement 1 with a diagram similar to Figure 18c and explained that the statement was true because 93% of the sample values would produce an interval that would capture the actual value of the parameter. The next statement posed a problem for Brody. As with many of the participants, Brody had it firmly engrained that probability does not coincide with confidence intervals. Therefore, with this statement linking probabilities with confidence intervals, Brody's initial reaction was to say this statement was not correct. I asked Brody what parts of the equation represented random variables, to which Brody responded with the appropriate statistics: "$\bar{x}$ is this is a variable that is a statistic, it's based on whatever that random sample is. So $\bar{x}$ is dependent on the sample itself. And as it is s because it's the sample standard deviation." To continue this discussion, I asked Brody if there was a difference between a theoretical sample or a collected sample. Brody stated:

Brody: Yes, in that if you've already collected it, then the $\bar{x}$ is no long, like, the $\bar{x}$ is no longer a random or it would still be a random variable, but you know the value of it. Whereas with if you haven't done yet, obviously, you don't know the value.

KR: So is there a difference? Can we talk about probability in different ways and the two situations?

Brody: I think in terms of you could you could look at probability differently on what is the probability of obtaining a sample monthly rent. Right. But I

> don't think you could say, I don't think it changes the probability for the actual mean monthly rent.

Brody has conceded that there is a difference between a situation when the sample has and has not been collected. He recognized that there was a difference in terms of probability but was still concerned with how the probability related to the [actualized] sample and [actualized] interval. The next question I asked Brody was Task 2, Jamie’s Colleague (see: Appendix). In Brody’s response to this question, he started to make the connection between the randomness, and thus probability, associated with the estimator and the fixed situation of the estimate. After thinking about the question, Brody stated that: “It's actually kind of makes sense. And I don't know if it's just because the way I was taught, I think, is that like, instinctively you go, it's like, oh, like it has to be 0 or 1.” He continued by demonstrating that if you have collected a sample, the actual value of the parameter is either in the interval or it is not. Brody stated:

> So, I guess it would be true that it's not until you set those parameters [Brody makes a gesture to symbolize the bounds of the interval, not parameters in terms of statistical parameters] that it either is or it isn't. … But if we haven't taken our sample yet, there is still a chance that it could be or that could not be and so I think I think your colleagues right. I don't know. This is the most in depth of actually explored this, I haven't really thought about it. … Because I've always been, I've always been taught that like, it's not a probability because it's set already and that like that's true. … But when it's not set, I think this is I think this is correct.

As Brody was acknowledging the connection between the randomness associated with and without an [actualized] interval, I referred Brody to a diagram he drew earlier in the interval (Figure 18e):

> KR: But 95% of your data is in here, or your, your statistics are in here.
>
> Brody: Yeah. And If that's so, if 95% of these are in this range, that means 95 percent would give me intervals that include that true, the true parameter. And so since this is the sampling distribution for a sample mean of size, whatever we whatever we're using, there's a 95% chance that a random value chosen from this distribution would be on the interval. And so yeah, there is a 95% chance that my sample include that my sample yields a confidence interval. That includes the true proportion, or true mean… But we can only say that before we do the sample.

At the conclusion of this part of the interview, Brody had connected the isolated knowledge about the long-run interpretation of the confidence level, the coverage probability, and the randomness associated with the difference between the estimator and the estimate. As a further validation of Brody's conceptualization of a confidence interval, during Interview 3, Task 1 (see: Appendix), Brody was the only participant to select Figure 14c as the best visualization of a confidence interval. During this question, Brody stated that because the value of the parameter is unknown, the graphs of the normal distribution centered at the statistic and the unknown value of the parameter were both not the correct visualizations of the confidence interval:

> this [pointing to Figure 14a, normal distribution centered at the statistic] maps what our actual interval is, but I - the - like sampling distribution doesn't have this

> normal curve cuz this is centered at the sample proportion instead of the population proportion, which again, we don't know. So, we don't really know where the histogram is centered. Therefore, like we can't really visualize it except as p. But the issue with this [pointing to Figure 14b, normal distribution centered at the unknown value of the parameter] one is that it's assuming that p is at the center of that interval. So neither of these [Figure 14a and b] really fits.

Instead, Brody selected the figure with the two normal distributions:

> This one [pointing to Figure 14c], I feel like does a better job of visualizing that. The interval is like not necessarily based on a histogram, necessarily. It's based solely on the sample proportion. And granted like the margin of error. We'll have will be based on what the margin like what the z-star or whatever this histogram will be.

He did, however, comment on the issue he brought up on Interview 2, the problem with estimating the standard error and concludes the picture may be misleading.

Gabe

Like Emma, Gabe has developed simultaneous conceptualizations of confidence intervals that can sometimes be in conflict with one another. During the first interview, Gabe interpreted the confidence interval using the correct interpretation [written]: "We are 95% confident that the true proportion of songs in the collab. playlist that belongs to Jamie is between .54 and .79." When asked to define the word confident, Gabe produce a secondary interpretation that stated that "95% of the possible proportions are within there [the interval]." Although it is not clear to which proportions Gabe is referring, multiple population proportions or multiple sample proportions or envisioning that the population data is composed of multiple proportions, this statement contradicts the statement Gabe originally wrote down. He continued these simultaneous conceptions of the interpretation of the confidence interval by stating a third interpretation: “between 54% and 79% of the samples taken from this playlist are going to have a high proportion of songs from Jamie rather than Alex.” Gabe also gave several definitions for the word confident over the course of the interviews. During the first interview, Gabe described the word confident as “how strongly you believe in this interval” and implied, during the second interview, that confident is strength in the sample collected. He stated that confidence comes from the confidence level (and thus, confidence coefficient) and the sample size. While discussing the difference between probability (and chance) and confidence, Gabe used the explanation that the confidence interval either captures or does not capture the actual value of the parameter as evidence the interpretation of the confidence interval does not relate to probabilities. From the three statements about the interpretation of the confidence interval and his definitions of the word confidence, it is unclear what Gabe really envisions as the interpretation of the confidence interval. It does suggest, which is confirmed on Interview 2, that Gabe is able to correctly identify the interpretation of the confidence interval but has not necessarily made deep connections to curricular concepts to internalize the meaning of this interpretation.

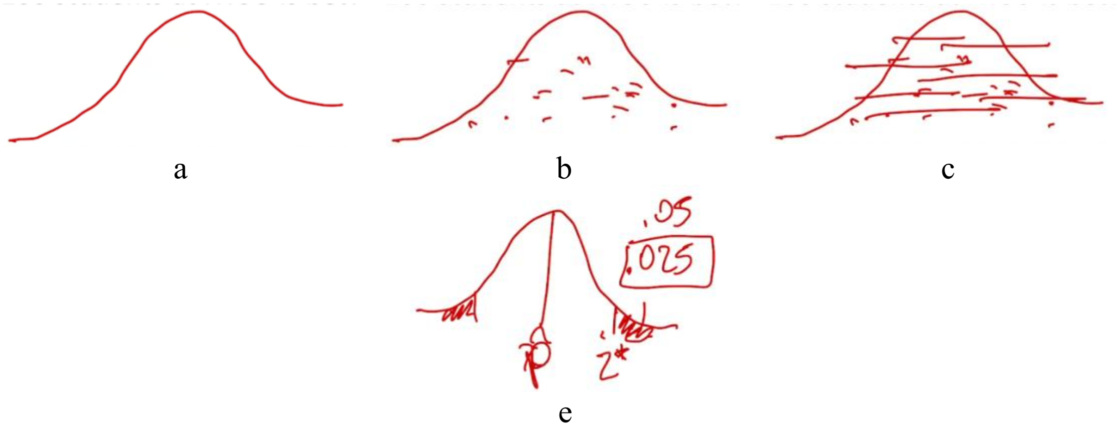
Similar to his interpretations for the confidence interval, Gabe’s statements and meanings for the interpretation of the confidence level were contradictory. While he was able to correctly identify the statement on Interview 2, Task 1, Question 3 that was the long-run interpretation of the confidence level, Gabe’s understanding of this interpretation of the confidence level was not necessarily reflected in his earlier statements about the confidence level. On interview 1, Gabe implied that the confidence level was the proportion of the time the [actualized] interval would be replicated or would present results that would be within the [actualized] interval. These statements coincide with the imagery that Gabe used to describe a statement that implied the confidence interval was the range in which 93% of repeated sample statistics would lie. In the explanation, Gabe stated that 93% of the samples would be within the [actualized] interval that was provided in the problem:

> What it means by 93% confidence is if we like, [drew Figure 19a] … If we found sample means from a population, the sample means might be like all over the place. [added dots seen in Figure 19b] I mean, if we do the same confidence interval, we think that 93% of these [added lines seen in Figure 19c] are going to be falling somewhere in this range [indicating the final figure in Figure 19c]. …
>
> Or uh 93% of these are going to include some portion of the range 705 to 793.

In this example, Gabe appeared to visualize the sampling distribution in such a way that the interval provided by the current sample would contain the 93% of the sample statistics. This would be the interpretation of the visualization in Figure 14a, a sampling distribution centered at the statistic. Gabe's initial drawing of a sampling distribution was this picture, a sampling distribution centered at the statistic (see: Figure 19e). As I directed his attention to the drawing, Gabe corrected the picture by stating the distribution was centered at the unknown value of the parameter. It is not clear, however, from his explanations that Gabe has connected the ideas of the [actualized] interval with the theoretical sampling distribution centered at the unknown value of the parameter (seen in Figure 14b). Instead, this conflicting image of a computed (actualized) interval and the theoretical sampling distribution appears to have created this doubt within Gabe that enables him to interpret the statement about the range where 93% of the repeated sample statistics would lie to be correct.

Figure 19

Gabe's Drawings during his interviews



Graduate Students

There were three graduate statistics students that were analyzed for this study: Joel, Liam, and Jace. Their conceptualizations of the interpretations of the confidence interval and interpretations of confidence level are provided below.

Joel

As a graduate student in statistics, Joel's theoretical knowledge is evident in both interviews. The theoretical knowledge, however, does not translate to deep conceptual knowledge or the ability to interpret the concepts of confidence interval or confidence level. During the first interview, Joel interpreted the confidence interval using a probabilistic statement that almost connects to a long-run interpretation of a confidence level: "to interpret you just like what I said the true ratio, the true proportion is like on average 95 time in this confidence interval." During Interview 2, Joel did select the correct interpretation of a confidence interval as being a true statement. It is unclear if "on average 95 times" is technically equivalent to the definition of confidence in the correct interpretation. It depends on whether or not Joel means "this confidence interval" as a random interval or as an actualized interval. If it is a random interval, then Joel's interpretation would be correct. If Joel intended "*this confidence interval*" to mean the actualized interval, then this statement is incorrect. This subtle difference can be seen in his explanation of the word confident. Initially, Joel stated that the meaning of the word confident confused him in both Chinese and in English. He said that he had used a video that explained the long-run interpretation of the confidence level as an explanation of the word confident. The video, Joel explained, demonstrated the confidence level by generating samples to see if the confidence interval captured the actual value of the parameter.

The programming language interpretation of the confidence level bled into other interpretations and explanations of the confidence interval and confidence level. Many of Joel's explanations relied heavily on "textbook definitions" or statistical theory but were often lacking the conceptual connections that were evident in Brody's explanations. For instance, Joel's second definition of the word confident was ultimately the same definition, but this time he implied this was a textbook definition:

> the formal one and because it comes with well-defined definitions, because we do mathematical worldly to statistics, after giving a concept we have a formal definition. So, what does this word mean? What does this concept mean? It comes up is a well-defined mathematical language, right? So, the here confidence we can

> interpret it as the like, what has that but here the sure probability and chance doesn't have certain definitions, or I just out of outside my knowledge

Of particular interest was the explanations that Joel provided for his answers to the correctness of the following statements:

1. The process used to generate confidence intervals will capture the actual mean monthly rent for students at HTSU approximately 93% of the time.
2. There is a 93% probability that the actual mean monthly rent for students at HTSU is within the interval $\bar{x} \pm \left(t_{n-1}^{*} \frac{s}{\sqrt{n}}\right)$.

Joel disagreed with statement 1 because of the "approximately 93% of the time." This statement was closest to Joel's previously stated interpretations of the confidence interval and confidence level. Joel did, however, agree with the statement 2. Other participants quickly dismissed this statement because of its use of the word probability, but Joel agreed with the statement saying: "yeah, because the confidence interval is kind of random itself. So, there the probability that this confidence interval capture, capture the true value." This statement Joel was referring to is actually a correct statement, but it should contradict Joel's stated interpretations and meanings. Instead, Joel should agree with the previous statement, but he does not. Statement 2 is more mathematical than statement 1, which potentially confirms that Joel's understanding of statistics is theoretically strong but in a procedural way. The conflict between mathematical-statistical theory and practical implications and interpretations can be seen in Joel's response to Jamie's Colleague. Based on the confirmation of statement 2, Joel should believe that Jamie's Colleague is correct in his statements. Joel, however, disagreed with the probability statement the colleague made (that prior to collecting the data, there is a

93% probability the interval captures the actual value of the parameter). Joel was pushed to explain the effect the situations when a sample has and has not been collected has on the calculation of probability. Using pivotal quantities and theoretical statistics notation, Joel eventually partially agreed with the colleague, stating:

> I think your statement is partially true because it just after collecting the sample that, like this probability or is just either zero or one, but why we still say like, we always say it's cannot be 0 or 1 we, instead of that we provide we provide this significance level, right, because we want to generalize our results to the underlying distribution.

Joel conceded that the second statement of Jamie's Colleague was correct, the probability is either 0 or 1, but was not able to envision the probability statement. Instead, he evoked the significance level, which Joel equated to a measure of accuracy. He used the significance level to reiterate the long-run interpretation of the confidence level. During the first interview, Joel's interpretation of the confidence level was a procedural explanation of the relationship between the confidence level, the margin of error, and the length of the confidence interval. During the second interview, Joel eventually disagreed with the correct interpretation of the confidence level, "approximately 93% of all samples of size 100 from the HTSU student body will produce 93% confidence intervals that capture the actual mean monthly rent for all students at HTSU," because of the first "93%." Joel instead stated that 100% of the confidence intervals will produce confidence intervals that capture the actual value of the parameter. It is not clear why Joel changed the two percentages, but the interpretation he provided for the confidence interval should have carried over to the interpretation of the confidence level.

During the course of the two interviews, Joel used distribution curves to explain the confidence coefficient and margin of error. He did not use the visualizations to explain the relationship between the actualized interval and the theoretical sampling distribution. During the third interview, Joel selected Figure 14a for introductory students to visualize a confidence interval, but Figure 14b to visualize the margin of error required to calculate the confidence interval. This may explain why Joel did not have a conceptualization of the confidence level as the coverage probability beyond the statement: $P\left(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\right) = 1 - \alpha$.

Liam

Out of the three graduate students analyzed for this study, Liam was the only one who has, as part of his assistantship, been a teaching assistant for the laboratory sections that supplement the lecture setting for the introductory statistics courses. Therefore, Liam has both teaching experience, which is evident in some of his explanations, and has taught from a conceptually driven curriculum that was, in part, designed by the current author. Liam described the interpretation of the confidence level procedurally, describing within his written statement the fill-in-the-blank parts:

> So you have to discuss what the interval is capturing. So you want to capture the true proportion of songs added by Jamie to the playlist. And since we didn't include the entire distribution in our creation of the confidence interval, then we can't be exactly certain that we capture the that true proportion. So that's why we have to add this qualifier that we're 95% confident, and our confidence interval that we created, and then you have to say what is the interval, that is between .52 and .79.

His explanation of the terms in the sentence, however, indicate that Liam has not necessarily developed a conceptual understanding of this interpretation. Liam's definition of the term confident simply was "doing what it's supposed to." This statement is not necessarily incorrect, but Liam was not necessarily able to describe this statement beyond "capturing the true proportion between these two values." Initially, Liam does not remember why the word confident and probability cannot be switched in the interpretation of the actualized interval. He eventually recalled that probability was not associated with the [actualized] interval because the [actualized] interval either captures or does not capture the actual value of the parameter. During the second interview, Liam was adamant that probability cannot be used to describe a confidence interval but stated that "sure" was acceptable because it is similar to "*93% belief*." This is potentially problematic if an individual sees probability and belief as equivalent statements. Although, there is no evidence in Liam's interviews that he either has separate definitions or similar definitions for belief and probability. As Liam discussed Jamie's Colleague (Interview 2, Task 2, see: Appendix), Liam did agree with both statements but admits that they sound contradictory, having particular trouble with the 93% probability associated with the random interval.

Liam's conceptualization of the interpretation of the confidence level is two part: 1) the long-run interpretation and 2) coverage probability. Like Brody, Liam is able to both interpret the confidence level using the long-run interpretation: "assuming we take samples and create 95% confidence intervals and take a whole bunch of those samples and a whole bunch of those intervals, then. In about 95% of those conference intervals, we will capture the true proportion." He drew the confidence interval applet (see: Figure

15) as a visualization of the confidence level (see: Figure 20a). He was able to connect this interpretation to the coverage probability, that the confidence-level% of statistics are within the confidence coefficient on the true sampling distribution. Liam explained the construction of the confidence interval using the conceptually based instruction in the laboratory manual:

> We know that roughly 95%, the distribution will be within two standard errors of the center at the true parameter, the true proportion, p [drawing the lines connecting -2SE and +2SE in Figure 20b]. So then, what we have to do is figure out how to reconstruct p from the - these two end points here [pointing to -2SE and +2SE in Figure 20b]. So, you would assume first that your lower edge came from this lower edge of the distribution [drew x on -2SE in Figure 20c]. And so we have a $\hat{p}$ here as equal to p minus the 2 standard errors [wrote $\hat{p} = p - 2SE$ in Figure 20c]. So that if we reconstruct that, that gives us p $\hat{p}$ plus two standard errors [wrote $\Rightarrow p = \hat{p} + 2SE$ in Figure 20c], and then assume over here [drew x on +2SE in Figure 20d], we have a different $\hat{p}$ this time, it's p plus two standard errors [wrote $\hat{p} = p + 2SE$ in Figure 20d]. And it's ah some algebra gives us $\hat{p}$ minus two standard errors [wrote $\Rightarrow p = \hat{p} - 2SE$ in Figure 20d].

Since this was the dictated laboratory activity, I asked if the activity made sense to him. Liam stated that it was not how he learned to calculate a confidence interval and expressed uncertainty about whether or not the activity was beneficial:

> It took getting used to that it does now. Like maybe after one full semester of teaching it clicks. It just seemed like I can write all this out and see how it gets us to what we want. And just seemed like, extra effort, I guess, when you could

> explain how I did it earlier about kind of giving, like a window around the area around the point estimate. Whereas this is more, I guess, concrete. And can gives the picture and some more principle-based understanding, like flow to reach the conclusion, instead of saying oh here's the formula and go work with it. So, I think it has its benefits. I just I'm not sure if learning this initially is better, or learning the formula and then explaining this is how we produce the formula.
>
> Like, I don't know which path gets you there. With more understanding.

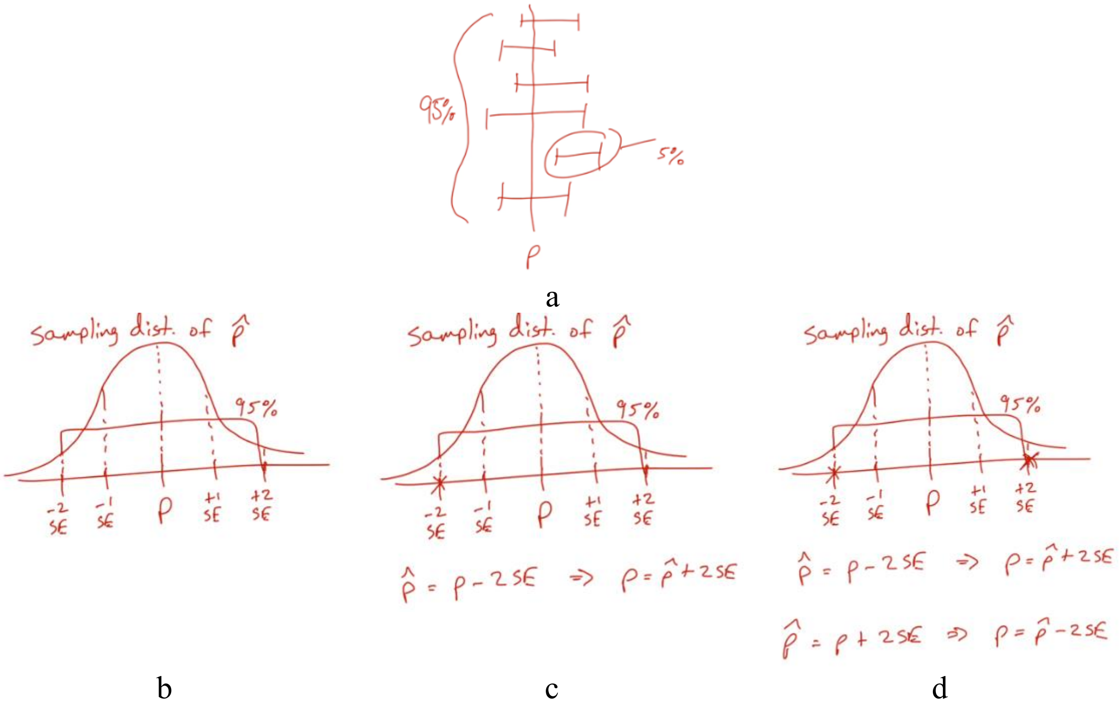
I include this quote for two main reasons: 1) to make the claim that while this was Liam's demonstrated knowledge, it may not, in fact, be his conceptual understanding, and 2) to demonstrate Liam's beliefs about procedural versus conceptual teaching. In fact, Liam admitted that when he envisions the calculation of a confidence interval, he manipulates the likelihood function and pictures the generic distribution shape to determine the confidence coefficients:

> That's a little different. So, you have some kind of like pivotal quantity between two quantiles, and then you kind of just shuffle stuff around until you get the parameter between two quantities. Which are, I guess, the end points of the interval? So, I don't really like picture. The distribution. When I'm creating the main time, I like picture, the distribution is when I'm thinking about which like, quantile is like in the tail region, which ones in like the fat part for like, not symmetric distributions. Yeah, my first instinct is not to picture like a visualization of the distribution.

The conflict between how Liam personally visualizes the construction of the confidence interval and coverage probability and how Liam teaches confidence intervals could account for the different, and sometimes conflicting, statements Liam made during the course of the interview.

Figure 20

Liam's Drawings From Interview 1


95%
5%
p
a
Sampling dist. of $\hat{p}$
95%
-2 SE
-1 SE
p
+1 SE
+2 SE
b
Sampling dist. of $\hat{p}$
95%
-2 SE
-1 SE
p
+1 SE
+2 SE
$\hat{p} = p - 2SE \Rightarrow p = \hat{p} + 2SE$
c
Sampling dist. of $\hat{p}$
95%
-2 SE
-1 SE
p
+1 SE
+2 SE
$\hat{p} = p - 2SE \Rightarrow p = \hat{p} + 2SE$
$\hat{p} = p + 2SE \Rightarrow p = \hat{p} - 2SE$
d

Jace

The general theme to Jace's interpretations of the confidence interval and the confidence level was repeated sampling. Over the course of the two interviews, Jace never interpreted an actualized interval. Instead, Jace focused on the long-run interpretation of the confidence level and repeated sampling. This reliance of repeated sampling was evident in his explanation of the calculation of a confidence interval. First, he explained that the best method to estimate the proportion of songs on the playlist that belonged to Jamie was through repeated sampling of the playlist. Jace stated that he was unable to calculate the confidence interval for the proportion of songs on the playlist that belonged to Jamie because there was only one sample and no variance. He interpreted the hypothetical confidence interval by referencing the capture / does not capture scenario and stating: “It's not 95% chance to be the true parameters. That 95% confidence interval means like, you do these experiment, like thousand times or 100 times, there are 95% chance that the proportion of the confidence interval contains the true parameters.” Jace further explains that there is a proportion of the time results would be within the [actualized] interval: “if we do the experiment 100 times, they are might be 95 times the estimated proportion would within .54 and .79 that range.” Despite not using the term confidence in his interpretation, I asked him to define the word confident, which likewise was interpreted using the long-run interpretation of the confidence level: “if you do 1000 or 100 times they are 95% times the experiment, the true parameters would be would drop in the would fall into the confidence interval.” When interpreting the confidence level, Jace focused on the width of the confidence interval’s relationship to the confidence level and the long-run interpretation of the confidence interval. During the second interview, Jace expresses that the words confident, sure, probability, and chance are all equivalent terms. Despite being clear during his explanation of the word confident during the first interview that there was no probability associated with the [actualized] interval, Jace stated that there was probability with the [actualized] interval capturing the actual value of the parameter: “confidence interval go back to the definition of the confidence interval. you do multiple tests there might be 95% chance the interval includes the true, like the actual mean think the 93% chances here.”

It is unclear if there were English to Chinese translation issues during some of the interviews. In particular, Jace had a problem with the interpretations provided during

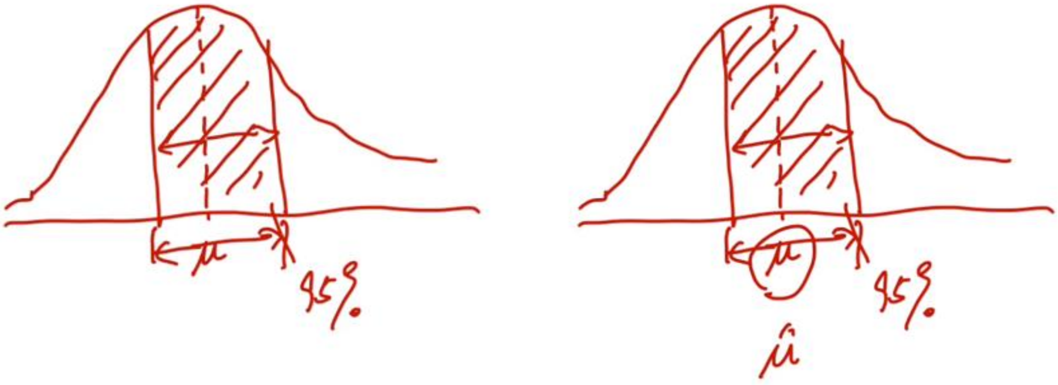
Interview 2, specifically with respect to the different phrases: capture, is between, and within. As a result, it is difficult to determine if Jace's word choices during conversations were great representations of his interval conceptualizations. Jace explained the confusion by evoking a fixed actual value of the parameter and fixed interval. He stated that for the confidence level to be interpreted, the actual value of the parameter would need to change to "capture" the actual value of the parameter:

> This sentence the weirdness of the sentence, like, what do you do is to capture the actual mean. It's like you're catching the true mean. But, it should be the actual mean falls into the confidence interval. I don't know if it is the same sentence, I mean the sentence meaning in English but to me it's kind of weird. It should be the actual mean falls into the confidence interval. … if I go to capture the actually mean is like in order to get the 93% of confidence interval I need to change the actual mean

The confusion of the action on the part of the actual value of the parameter is continued when Jace discussed his visualization of a confidence interval. He began by explain that the distribution would be normal, but stated that the confidence interval is the portion of the distribution within which 95% probabilities lies, centered around the unknown value of the parameter (see Figure 21a). Jace continued by stating that since the value parameter, μ, was unknown, it can be substituted with $\hat{\mu}$. As I pressed to determine what Jace pictured as being within the 95% region, Jace stated that the 95% was connected to the actual value of the parameter. Again, Jace reiterated that the region contains the range within which the actual value of the parameter would fall 95% of the time: "this is the confidence we calculated it and if we do the experiment 100 times, they are might be 95 times the estimated proportion would within .54 and .79 that range." It may be relevant to state that the repeated nature, and the presence of probability on the value of the parameter, was more consistent with Bayesian perspective of probability and inference. Therefore, it may be unclear without further interviews to determine if Jace truly has the conception that the actual value of the parameter "falls" within the diagrams he drew, if it was an English-to-Chinese translation issue, or a Bayesian versus Frequentist differences.

Figure 21

Jace's Drawings on Interview 1



a

b

APPENDIX M

CLASSIFICATIONS OF CONCEPTIONS OF THE INTERPRETATIONS

Table 19

Categorization of Participants' Conceptualizations of Confidence and Confidence Level by Statistical Course

		Intro	Intermediate			Capstone				Graduate			Total
		Tiana	Kiara	Aiden	Logan	Diana	Emma	Brody	Gabe	Joel	Liam	Jace	
Confidence	Quantifiable Measure					Yes		Yes	Yes	Yes	Yes		5
	Sureness/ Belief	Yes	Yes	Yes					Yes		Yes		5
	Confusion: Conf/Chance	Yes	Yes	Yes						Yes	Yes		5
	Equival: Conf/Chance				Yes		Yes					Yes	3
Confidence Level	Confidence Coefficient	Yes	Yes			Yes							3
	Relation to Width		Yes	Yes	Yes		Yes			Yes		Yes	6
	Probability/ Accuracy	Yes		Yes			Yes		Yes	Yes		Yes	6
	Coverage Probability						Yes	Yes	Yes		Yes		4

Table 20

Classification of Participants' Conceptualizations of the Interpretation of Confidence Interval

		Correct				Not Probability		Replicate of Experiment		Incompatible		
		Brody	Liam	Emma	Gabe	Diana	Jace	Logan	Joel	Tiana	Kiara	Aiden
Confidence Interval	Correct	Yes	Yes	Yes	Yes							Yes
	Capture/ Not Capture	Yes	Yes	Yes	Yes	Yes	Yes					
	Long-run/ Probability					Yes	Yes	Yes	Yes			Yes
	Confounding									Yes	Yes	

Table 21

Classification of Participants' Conceptualization of the Interpretation of Confidence Level

		Correct			Probability and Width			Parallel		Percentage		
		Brody	Liam	Diana	Kiara	Logan	Joel	Emma	Jace	Aiden	Gabe	Tiana
Confidence Level	Long-run	Yes	Yes	Yes				Yes	Yes			Yes
	Coverage Probability	Yes	Yes		Yes	Yes	Yes	Yes	Yes			
	Percentage				Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 22

Classification of Participants' Conceptualizations of Confident

		Quant Measure		Belief		Towards Probability				Probability		
		Brody	Diana	Gabe	Liam	Tiana	Kiara	Aiden	Joel	Logan	Emma	Jace
Confidence	Quantifiable Measure	Yes	Yes	Yes	Yes				Yes			
	Sureness/ Belief			Yes	Yes	Yes	Yes	Yes				
	Confusion: Conf/Chance				Yes	Yes	Yes	Yes	Yes			
	Equival: Conf/Chance									Yes	Yes	Yes

Table 23

Classification of Participants' Conceptualizations of Confidence Level

		Coverage Probability		Developing						Surface Level		
		Brody	Liam	Jace	Joel	Emma	Gabe	Aiden	Logan	Tiana	Kiara	Diana
Confidence Level	Confidence Coefficient			Yes	Yes			Yes		Yes	Yes	Yes
	Relation to Width			Yes	Yes	Yes		Yes	Yes		Yes	
	Probability/ Accuracy					Yes	Yes			Yes		
	Coverage Probability	Yes	Yes			Yes	Yes					

APPENDIX N

CLASSIFICATION OF THE CONCEPT OF CONFIDENCE INTERVALS

Table 24

Participants' Conceptualization Themes for Concept of Confidence Intervals by Statistical Course

			Intro	Intermediate			Capstone				Graduate		
	Category	Sub-Theme	Tiana	Kiara	Aiden	Logan	Diana	Emma	Brody	Gabe	Joel	Liam	Jace
Interpret CI	Correct	Correct			🟥			🟥	🟥	🟥		🟥	
		Hypothesis						🟫		🟫			
		Formulaic							🟫			🟫	
	Capture/Not Capture						🟥	🟥	🟥	🟥		🟥	🟥
	Long-run and Probability				⬛	⬛	🟫				🟫		🟫
	Confounding Ideas		🟫	🟫									
Interpret CL	Long-run	Correct					🟥	🟥	🟥			🟥	
		Repeated											🟥
		Procedural	🟫										
	Coverage Probability	Connected							🟥			🟥	
		Width		🟫				🟫					
		Compromise				🟫					🟫		🟫
	Percentage	Value	🟫	🟫	🟫			🟫					
		Actualized				⬛		⬛		⬛	⬛		⬛
Confident	Quantifiable Measure	Estimating					🟫		🟫	🟫		🟫	
		Process				🟫			🟫	🟫		🟫	
	Sureness/ Belief		🟫	🟫	🟫					🟫		🟫	
	Confusion of Confidence		🟫	🟫	🟫						🟫	🟫	
	Equivalent of Confidence					⬛		⬛					⬛
Confidence Level	Confidence Coefficient		🟫	🟫			🟫						
	Relation to Width			🟫	🟫	🟫		🟫			🟫		🟫
	Probability and Accuracy		🟫		⬛			🟫		⬛	🟫		⬛
	Coverage Probability							🟥	🟥	🟥		🟥	

Note. Red indicates productive and correct reasoning. Tan indicates developing reasoning. Black indicates incorrect reasoning.