A Multi-Modal Multiple Descriptive Case Study of Graduate Students' Statistical Thinking in Statistical Tests Seven Months After Completing a Simulation-Based Introductory Level Course

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

V.N. Vimal Rao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Robert delMas, Dr. Andrew Zieffler (advisors)

May 2023

Acknowledgements

The path from matriculation to dissertation is one I never walked alone. First and foremost, thank you to my family – my mother, my father, my sister, my grandparents, my uncles and aunts and cousins – for their unwavering love and support.

Thank you to my statistics education family for accepting me with open arms and unconditionally supporting me in my academic ventures. Thank you to my advisors, Robert delMas and Andrew Zieffler, who patiently guided me on my journey, always nudging me in the right direction while letting me find my own path. Thank you to Karen Schlumpf, Timothy McCall, Ann Brearly, and Laura Le, for their mentorship in my development as an instructor. Thank you to my statistics education siblings nationwide – Nina Bailey, Sayali Phadke, Elijah Meyer, Charlotte Bolch, and Kit Clement – for providing camaraderie as we completed this journey together. Thank you to my elder UMN statistics education siblings – Chelsey Legacy, Jonathan Brown, Ethan Brown, Elizabeth Fry, Laura Le, Matthew Beckman, Nicola Justice, Laura Ziegler, and Anelise Sabbag – for always providing sage guidance as I follow in their footsteps. I want to especially thank my Stat Ed sibling Nina Bailey for sharing in the special adventure of writing a dissertation. I also want to especially thank my UMN Stat Ed siblings Jonathan Brown, for taking me under his wing as a new student in the program, Regina Lisinker and Pablo Vivas Corrales, for reminding me about the joys of inquiry and discovery, and Chelsey Legacy, who patiently entertained all of my ideas, and with whom I have walked this path from my first day.

Thank you to my UMN family, for always providing a sympathetic ear to support me through the rollercoaster of emotions that one faces in graduate school. Thank you to Carlos Chávez, Rina Harsch, and Ashley Hufnagle for always providing a reason to laugh

and smile. Thank you to Vijay Marupudi, Reba Koenen, Shelby Weisen, Megan Goeke, Maggie Sullivan, Ali Fulscher, Tai Do, Corissa Rohloff, Rik Lamm, and Jesslyn Valerie, for always reminding me that it's the people that you spend time with that is most important in life. Thank you to Justin Baker and Ashley Hufnagle for always encouraging me to dream big. My deepest thanks to Kristin Running, for sharing in every celebration and always believing in me. Thank you to Sashank Varma, Jeffrey Bye, and David DeLiema, for seeing potential in my ideas and adopting me as a mentee. I would not be the person I am today without each of your love and support.

This dissertation would not be the same without the contributions and support of several individuals. Thank you to my advisors, Robert delMas and Andrew Zieffler, for their helpful comments on every single iteration of this project. Thank you to my committee members, Panayiota Kendeou and Erin Baldinger, for pushing me in my thinking and providing helpful resources for the design and analysis of this study. Thank you to Martin Van Boekel, Michael Harwell, Kristin Running, Chelsey Legacy, Nina Bailey, Jimin Park, Kelsey Will, Valerie Barbaro, and Ethan Brown for guiding me in my conceptualization of this study and for their feedback on various iterations of this work. Thank you to David DeLiema and Zachary Carpenter for their consultation and advice in the design of this study. Thank you to Karen Givvin, Ji Son, Jim Stigler, and the members of UCLA's Teaching and Learning Lab, for their helpful comments on the analysis, interpretation, and communication of the data. Special thanks to Zachary Carpenter, without whom I would not have been able to collect gaze data.

And finally, thank you to the six graduate students who spent time with me to help me understand how they think. Without you, there is no study.

Invocation

ॐ

सह नाववतु । सह नौ भुनक्तु । सह वीर्यं करवावहै ।

तेजस्विनावधीतमस्तु मा विद्विषावहै ।

ॐ शान्तिः शान्तिः शान्तिः ॥

Om

saha naavavathu. saha nau bhunakthu. saha veeryam karavaavahai.

theijasvi naavadheethamasthu maa vidvishaavahai.

Om shaanthih shaanthih shaanthihi.

Om

May we be blessed in our studies. May we be nourished with knowledge.

May we pursue our studies energetically and effectively.

May our studies be enlightening and without strife.

Let there always be peace in our bodies, minds, and spirits.

Dedication

To my parents, Vidya and Shyam, and their support and encouragement in all I do.

Abstract

Though statistical testing is commonly practiced, the logic of statistical tests is confusing, thinking about distributions is difficult, and the way statisticians formulate expectations as probability distributions is poorly understood. To support instruction, the statistics education community has increasingly utilized simulation-based pedagogies that place the logic of statistical inference at the core of instruction. Might this approach support and sustain the development of graduate students' statistical thinking, especially during statistical testing? How do graduate students, who have completed a simulation-based course, think while conducting statistical tests, months after completing the course?

To answer these questions, a multi-modal multiple descriptive case study of six graduate students in the educational sciences was conducted. Data sources included audio, video, and gaze recordings, analytic memos generated by the researcher, as well as written artifacts generated by the participants. Participants generated concept maps for the logic of statistical tests, conducted statistical tests using statistical software, interpreted results from statistical tests, and participated in a retrospective video-cued interview. Data were analyzed through an interpretivist epistemological stance and employed the constant comparative method to identify relevant moments across all data artifacts to credibly describe participants' thinking.

Results suggest that students' planning (i.e., deciding what to do and when to do it) was generally quite good. However, students generally struggled in monitoring and evaluating their plan (i.e., ensuring that the plan was being executed correctly, and that no changes to the plan were needed). Furthermore, they generally did not seem to think about null models, core to the logic of statistical testing. Instead, they focused on point and

interval estimates for statistics of interest, and primarily thought about sampling variability in terms of a bootstrap dot plot, if at all.

This study is one of the first to examine graduate students' statistical thinking several months after the completion of a simulation-based introductory course. How students were thinking – generally able to reproduce a plan for analyzing the data consistent with what they were taught, and with a focus on variability through the examination of a bootstrap dot plot – suggests that statistics instructors might anchor instruction about statistical tests to descriptive statistics and their interpretation and contextualization. Furthermore, it suggests that the likelihood approach to statistical inference, evaluating hypotheses against given data, may be conceptually easier for students to think about.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Statistical tests, and their seemingly incomprehensible mathematics, reign with tyranny across the sciences (e.g., Lambdin, 2012). At least, this is how it used to be (Wasserstein et al., 2019).

With scholars decrying a bastardization of the classical statistical testing procedure that has promulgated throughout common practice (e.g., Cohen, 1990; Gigerenzer, 2004), many are turning away from statistical tests and their misunderstood null hypotheses and *p*-values. Some argue that we should shift to a different school of statistics, such as the Bayesian school (e.g., Kruschke & Liddell, 2018). Some argue that we should abandon testing in favor of estimation alone, prioritizing confidence intervals and estimates of effect sizes (e.g., Cumming, 2014).

Yet, none of these proposals resolve the issues at the heart of the matter – the logic of statistical tests is confusing (delMas, 2004), the manner in which statisticians formulate their expectations as a probability distribution is poorly understood (Nickerson, 2000), and more generally, reasoning about distributions is difficult (e.g., Reading & Reid, 2006).

These problems are not new – there has always been confusion in the application of statistical testing methods (e.g., Boring, 1919). Over the years, statistics educators have tried many ways to improve their teaching and communication of statistical theory and practice, and starting in the 1980s, they began to utilize simulation-based pedagogies.

Simulation-based inference (SBI) as it became to be known continued to slowly blossom into full-fledged SBI curricula. The statistics education community generally believed that this pedagogy was more apt to developing students' conceptual understanding of statistics, by placing the logic of statistical inference at the core of instruction and

eschewing units on probability that was mathematically challenging for many students (Cobb, 2007).

By 2022, evidence from a small yet growing body of research generally supported the claim that SBI curricula were at least no worse than traditional parametric-based introductory curricula at developing students' conceptual understanding (Brown, 2019). While SBI curricula had been predominantly utilized to teach undergraduate students, some master's levels courses were also taught using SBI pedagogies (e.g., Brown, 2021).

Here then is a potential amelioration for problems with statistical testing in scientific practice – SBI, a recommended pedagogical approach to developing students' statistical thinking, may be a useful tool in training graduate students. As researchers applying statistical methods or practitioners interpreting statistical results, graduate students need to understand the logic of statistical inference (e.g., APA, 2017; GAISE, 2016). As science is fundamentally about theory generation and theory testing, graduate students need to be fluent in at least one mechanism of statistical testing, as statistical testing and inferential thinking are an integral aspect of scientific thinking (Dunbar & Fugelsang, 2005). Might SBI curricula be able to support graduate students' development of statistical thinking and an understanding of the core logic of statistical testing?

This is the central purpose of this dissertation, which considers the case of graduate students who have completed a master's level introductory statistics course utilizing an SBI curriculum. How do these graduate students think when they conduct statistical tests?

A thorough understanding of their thinking allows us to evaluate the potential benefits of an SBI curriculum for these students, identify conceptual difficulties that can

be addressed with pedagogical reform, and inform reformation in the practice of statistics to address extant controversies and crises.

To answer this question, Chapter 2 begins with a review of the logical foundations of statistical testing, and subsequently presents recent research on students' thinking within and after having completed an SBI introductory level statistics course. This background information sets the foundation for the main study conducted as part of this dissertation, the investigation of graduate students' thinking in statistical tests. Chapter 3 describes the study design and Chapter 4 reports the study results. Chapter 5 summarizes the study results in light of previous work about what is known about graduate students' thinking, and includes discussion of possible directions for future research, teaching, and practice.

Chapter 2: Background

What is statistical testing? What is the core logic of a statistical test? Does this change when statistical tests are conducted with SBI methods, instead of parametric-based methods? This chapter first presents a short history and philosophy of classical significance testing to answer the question “What is the core logic of significance testing?”. The chapter next briefly describes the philosophy and history of statistical thinking in statistical tests within the simulation-based approach to significance testing. Next, this chapter summarizes recent research focused on the following questions:

- To what extent do students’ difficulties, documented within parametric-based statistics curricula, persist even in simulation-based curricula?
- What is students’ understanding of null model simulators?
- What if any unique aspects to students’ thinking about significance testing emerge with simulation-based methods?

2.1 What is the core logic of statistical testing?

In early 2020, the World Health Organization (WHO) declared a global pandemic due to a novel coronavirus, COVID-19, and attention soon focused on experimental studies testing possible cures (WHO, 2020). In one such study, Wang et al. (2020) compared the median time to clinical recovery in an experimental group receiving a new treatment to a control group that did not – the median time was 21 days in the experimental group, two days less than the control group's median time of 23 days. While 21 days is shorter than 23 days, this difference could easily occur by a chance coincidence even if the drug had no effect whatsoever. Therefore, one should be hesitant to infer that the drug was effective at reducing recovery times (based on this evidence alone) since there exists a sufficiently plausible contradictory theory.

This type of reasoning based on a chance model can be traced back to the early 18th century (and perhaps may have been in use even earlier; Stigler, 2016). It is an informal version of a significance test and is an example of inferential reasoning (Zieffler et al., 2008). Inferential reasoning and statistical testing are integral aspects of scientific thinking and reasoning (Dunbar & Fugelsang, 2005).

Understanding and using the basic ideas of statistical inference is one of the key recommended learning goals for introductory level statistics courses at the post-secondary level (GAISE, 2016, Goal 7). Statistical inference refers to the act of using data from a sample to probabilistically describe a larger population (Makar & Rubin, 2009). The act of making inferences that extend beyond observed data is an inherently uncertain task, and is called the problem of induction (Henderson, 2020). While this uncertainty plagues all

forms of inference, statistical inference is based on utilizing probability to think and reason about this uncertainty (Romeijn, 2017).

Statistical inference can be divided into formal and informal variants. Formal statistical inference includes specific computations or formal tests (e.g., interval estimates, significance tests) while informal statistical inference refers to a broader set of concepts and understanding without requiring specific procedures (Tobías-Lara & Gómez-Blancarte, 2019). Statistical inference is often further sub-divided into the production of estimates and the testing of hypotheses (e.g., Wald, 1939). Estimation and testing might be characterized by distinct questions such as "What is the population like?" and "Is the population like X?" respectively.

Significance tests are one type of statistical test of a hypothesis, in which a claim about a population is epistemically judged (Moore et al., 2013). In the classical school of statistics, this is achieved by considering the expected consequences of a claim against observed evidence. The existence of a claim about the population in the significance testing procedure often presents difficulties for students, as they must reason about the hypothetical claim, the observed evidence, and the unknown truth about a population (Vallecillos & Batanero, 1996).

To establish a clear definition of the logic of significance testing, the next sections turn to philosophy and history to provide an answer. Specifically, it examines primary historical and philosophical sources that detail the origin of significance testing, before describing modern simulation-based curricula and how they attempt to develop students' understanding and thinking about significance tests.

2.1.1 Philosophy of Hypothesis Evaluation

Significance testing is fundamentally about comparing observed evidence to a hypothesis. The underlying philosophical theory detailing the relationship of evidence and hypotheses is called Confirmation Theory (Romeijn, 2017). Confirmation entails both positive affirmation of a hypothesis as well as disconfirmation, or a confirmation of a hypothesis's negation.

There are generally three approaches to confirmation discussed by 20th century philosophers: confirmation by instances, hypothetico-deductive confirmation, and Bayesian confirmation (Norton, 2005). Confirmation by instances was formalized in the early 20th century by Carl Hempel. Acknowledging the impossibility of induction to definitively prove universal truths, confirmation by instances instead focuses on the development of a hypothesis relative to observed evidence (Hempel, 1945). Observed evidence is said to Hempel-confirm a hypothesis if and only if the evidence can be considered an instance of, or is consistent with, the hypothesis. For example, observing an individual who is a human and has 10 fingers Hempel-confirms the hypothesis that 'All humans have 10 fingers', while observing an individual who is a human and does not have 10 fingers Hempel-disconfirms the hypothesis.

While the logic of confirmation by instances evaluates hypotheses on the basis of observed evidence, the hypothetico-deductive approach to confirmation reasons about observed evidence on the basis of a hypothesis. Observed evidence hypothetico-deductively confirms a hypothesis if and only if the hypothesis implies the evidence, or that the evidence is a consequence of the hypothesis (Sprenger, 2011). Similarly, observed evidence hypothetico-deductively disconfirms a hypothesis if and only if the hypothesis

implies the negation of the evidence, or that the negation of the evidence is a consequence of the hypothesis.

Returning to the previous example, the hypothesis ‘All humans have 10 fingers’ implies that an observed human must have 10 fingers, and thus observing such an individual hypothetico-deductively confirms the hypothesis. While the results in this case are the same as in confirmation by instances, the underlying logic is fundamentally different, as hypothetico-deductive confirmation requires postulating expected characteristics of observations on the basis of the hypothesis. Aside from these differences, both confirmation by instances and hypothetico-deductive confirmation depend on formal logic and suffer from many logical paradoxes. Furthermore, and perhaps most relevant to statistical inference, neither approach explicitly addresses the uncertainty inherent to inference (i.e., neither approach attributes probabilities to either observed evidence or hypotheses).

Bayesian confirmation, on the other hand, explicitly aims to incorporate probability into the process of confirmation, and does so by assigning probability values to hypotheses (Talbott, 2016). Bayesian confirmation is based on calculating the probability of a hypothesis (i.e., $P(H)$), as well as the probability of that hypothesis conditioned on having observed some evidence (i.e., $P(H|e)$). A hypothesis is Bayes-confirmed by this observed evidence if and only if the probability of the hypothesis conditioned on the evidence is greater than the initial probability of the hypothesis (i.e., $P(H|e)>P(H)$), while it is Bayes-disconfirmed by the evidence if the opposite is true (i.e., $P(H|e)<P(H)$). In other words, if in light of the observed evidence the probability of a hypothesis increases from its previous state, the evidence Bayes-confirms the hypothesis. Although Bayesian confirmation does

not suffer from the same logical paradoxes that plague confirmation by instances and hypothetico-deductive confirmation, it suffers from its own logical paradoxes. However, and most importantly when considering a philosophical basis for statistical testing, Bayesian confirmation is dependent upon choosing an interpretation of probability that allows probability to be assigned to statements and hypothesis.

There are generally four recognized philosophical interpretations of probability: relative frequency, propensity, logical, and subjective (Hájek, 2019). In the relative frequency interpretation, probability is defined as the limit of the relative frequency of a repeatable event (von Mises, 1939). This interpretation is the most restrictive interpretation of probability, as the relative frequency is undefined for non-repeatable events and is not used to assign probabilities to statements. For example, the probability that a fair coin lands head when flipped is well-defined, while the probability that a particular vase will break when dropped is not, as flipping a single coin is a repeatable event while dropping a single vase is not a repeatable event (after the first occurrence of it breaking).

The propensity interpretation accounts for this by expanding on the relative frequency interpretation to include the tendency or disposition of an event occurring (Popper, 1959). The propensity of an event becomes manifest as the event's relative frequency in the case of repeatable events, such as flipping a coin, but is still well defined for non-repeatable events, such as dropping the vase.

Both the relative frequency and propensity interpretations are considered physical probabilities, as they are physical characteristics of an event such as a ‘coin flip’ or a ‘vase drop'. In contrast, the logical and subjective interpretations are considered epistemic,

dealing with what individuals or communities of people believe or know, and are well-defined even when applied to statements.

In the logical interpretation, probabilities represent a degree of belief or credence in a statement for a community of rational persons with the same information (Keynes, 1921). For example, a group of paleoanthropologists might ascribe a probability of 0.85 to the Out of Africa Theory, a statement that modern humans originated from the African continent and subsequently migrated across the world. However, this interpretation does not leave room for dissent, and if two persons from the same community of rational persons with the same information assign different probabilities to a statement, at least one of them must be wrong. The subjective interpretation accounts for this by interpreting probability as the representation of an individual (rational) person's degree of belief that a statement is true (de Finetti, 1937). The subjective interpretation is thus the most liberal of the four interpretations of probability.

While probability in the logical and subjective interpretations are uniquely able to ascribe probabilities to hypotheses, in both interpretations, ascribing probabilities to events is also well-defined, and may even be equivalent to values that may be ascribed by the physical interpretations of probability, although not necessarily so. For example, within a physical interpretation of probability, the probability of a 'coin flip' resulting 'heads' may be .50. If an individual, or a group of individuals for a logical interpretation of probability, believes that the probability of the coin landing 'heads' is equal to .50, then all interpretations would assign the same probability value to the 'coin flip' event. In this way, philosophically speaking, the logical and subjective interpretations of probability can be utilized in all the same ways that the physical interpretations can be used, plus some.

As statistical inference utilizes probability to reason about the uncertainty of inferences, the choice between the physical interpretations of probability and the epistemic interpretations of probability is fundamentally intertwined with the different approaches to confirmation and their methods for reasoning about the relationship between observed evidence and a hypothetical claim. Classical statistics was heavily influenced by scholars who subscribed to a physical interpretation of probability (Romeijn, 2017). These scholars rejected the radical subjectivism advocated by scholars such as de Finetti, and with it, interpretations of probability that supported ascribing probabilities to candidate hypotheses.

Since physical interpretations only assign probabilities to events, these founding scholars of the classical approach to statistics necessarily adopted an approach akin to hypothetico-deductive confirmation, taking a hypothesis as given and reasoning about the consequences of that hypothesis in terms of the probabilities the hypothesis ascribed to all possible events. This choice required that observed evidence be treated as a single event from an infinite collection of events based on a repeatable process, thus allowing it to be ascribed a probability of occurring. A hypothesis could then be made about the probability of such events occurring, or the relative frequency of observing such an event in an infinite repetition of the evidence-generation process (e.g., Neyman, 1937). On this basis, an inference could be made about the hypothesis, and quantitatively described in terms of the quality of or support for the inference with notions such as significance, likelihood, or confidence.

In this manner, significance testing, one of the first approaches to formally evaluating statistical hypotheses, concerned itself with the probabilistic consequences of a

given hypothesis, and whether observed evidence deviated significantly from a hypothesis's probabilistically expected events.

2.1.2 Classical Statistical Testing

Expressions of probability to describe events long predate the formalization of modern statistical inference and significance testing. The first known formal probability calculation for an observed event under a candidate hypothesis hails from 1710, when John Arbuthnot analyzed the number of births by biological sex when examining birth records from the London area over 82 years (Stigler, 2016). Arbuthnot observed that in all 82 years there were more male births than female births, but first considered the plausibility of random chance producing the observed pattern before drawing conclusions. Arbuthnot hypothesized and assumed the chance of more male births in any one year was equal to 0.5, with the other potential outcome being more female births. Arbuthnot then specified a probability model to describe expected patterns produced by random variation based on this assumption. Using this model, Arbuthnot calculated that if the assumption was true, then the probability of all 82 out of 82 years having more male births was roughly equal to 0.02 septillionths (i.e., $2 * 10^{-25}$). With such a low probability, Arbuthnot concluded that the assumption must have been incorrect.

Two centuries later, such assumptions were codified as null hypotheses, their associated probability distributions as null models, and the probability calculations of the observed evidence as *p*-values (Fisher, 1925, 1935). While Fisher was not the first person to write about significance testing (e.g., Pearson, 1900), the popularity of Fisher's 1925 book *Statistical Methods for Research Workers* and 1935 book *The Design of Experiments* spread these concepts to a wide audience beyond the realm of statisticians, popularizing

their use (Stanley, 1966). Fisher (1935) defined the null hypothesis as the basis for the specification of a probability distribution, and that this probability distribution (i.e., the null model) would in turn serve as the basis for a significance test.

To Fisher, null hypotheses were a characteristic of all experiments, and experiments existed solely to give evidence a chance at disproving a null hypothesis. The experiment functioned as an infinitely repeatable event, often with a component random process, enabling a physical interpretation of probability to be assigned to possible outcomes based on the null hypothesis of the experiment. These probabilities could then be interpreted in a manner akin to a probabilistic hypothetico-deductive disconfirmation of the null hypothesis.

For a hypothesis to qualify as a null hypothesis, Fisher stipulated that it must be exact in its specification of a probability distribution which could subsequently be used to create an exact statistical criterion. The probability distribution represents "the frequencies with which the different results of our experiment shall occur" (Fisher, 1935, p. 190).

The statistical criterion, or significance level, demarcates the threshold beyond which observed evidence would, in relation to the null hypothesis, present a logical disjunction – either an extraordinary coincidence has occurred or the hypothesis is likely incorrect (Fisher, 1956). Observed evidence laying beyond the significance level thus constituted "rational grounds for the disbelief it engenders [in the null hypothesis]" (Fisher, 1956, p. 43).

Any hypothesis meeting the criteria for exactness could be chosen and given a chance to be disproved by this procedure. Thus, to Fisher, the null hypothesis was simply the hypothesis to be nullified via experimentation (Cohen, 1994). Fisher alternately

referred to the probability distribution specified by the null hypothesis as the "random sampling distribution on the null hypothesis" (Fisher, 1935, p. 62) and the "sampling distribution completely determined by the null hypothesis" (Fisher, 1935, p. 192). Today, we call this distribution the null model.

In this way, the existence of a null model is exactly the characteristic that makes a candidate hypothesis a null hypothesis. These null models thus provide the means to establish statistical criteria with which to compare observed data to the null hypothesis, all in the service of the potential nullification of a hypothesis through experimentation.

2.1.3 Alternate Approaches to Statistical Testing

A significance test taking a hypothesis as given and considering evidence that the hypothesis implies is only one approach to statistical testing. Recall that confirmation by instances takes the observed evidence as given and considers whether the evidence contributes to the development of a hypothesis. In this manner, one can consider the plausibility or credibility of potential values of a parameter given some observed sample statistic. Hempel called this an inductive-statistical explanation, while modern scholars consider plausibility by using what is known as a likelihood function (Barnard, 1967). This approach is an example of one of the major schools of thought for statistical inference, known as the Likelihood school. The Likelihood school differs from the classical school by emphasizing the likelihood function, and conducting both estimation and testing procedures by evaluating the entire likelihood function based on the observed evidence in a manner akin to abductive reasoning (Edwards, 1972).

Along with the classical school and the Likelihood school, the third of the three most popular schools of thought is the Bayesian school (Bandyopadhyay & Forster, 2010).

The Bayesian school fundamentally differs from both the Likelihood school and the classical school by relying on an epistemic definition of probability and assigning probabilities to hypotheses, through either a logical or subjective interpretation. The probability of the hypothesis prior to the collection of new evidence is called the prior probability, and the probability conditioned on the new observed evidence is called the posterior probability. In this manner, parameter values with posterior probabilities less than their prior probabilities are Bayes-disconfirmed.

Not only are there multiple approaches to the evaluation of a single hypothesis based on observed evidence across the three major schools, but each school of statistics also has a different approach to generating decision rules for selecting between competing hypotheses than their approach about epistemic judgements about evidence or hypotheses.

Within the classical school, the hypothesis testing approach is most common for generating decision rules, and seeks to minimize errors associated with incorrectly choosing one hypothesis over another. This is done by specifying decision criteria based on comparisons of the probability of observing events given each candidate hypothesis (Neyman & Pearson, 1928). In the Likelihood school, the likelihood ratio test is used to identify whether two models significantly differ by comparing their likelihood functions, while maximum likelihood estimation simply selects the most likely hypothesis on the basis of a likelihood function (Edwards, 1972). In the Bayesian school, the Bayes factor, the ratio of the posterior probability of two hypotheses, is used to select the hypothesis with the higher probability (Jeffreys, 1961).

A classical hypothesis test can be considered a before-data decision rule, specified without incorporating observed evidence, whereas the likelihood and Bayesian approaches

explicitly include the observed data in the selection of one of the competing hypotheses through the likelihood function or posterior probability distribution, which itself is partially based on the likelihood function (Hacking, 1965).

It is important to note that the task of selecting between competing hypotheses is fundamentally different from one in which the goal is to make an epistemic judgement about a single hypothesis. In these decision-based approaches to statistics, the goal is explicitly to provide a strategy for selecting one out of a set of competing hypotheses. Neyman characterized this method not as a theory of inference, but a theory of behavior (Neyman, 1952). Neyman rejected significance tests that only explicitly considered a single hypothesis, believing that researchers necessarily subconsciously consider an alternative hypothesis to be true if the single candidate hypothesis is rejected, and that it would be better to explicitly consider two candidate hypotheses. Epistemic judgements about these accepted hypotheses, (i.e., to what extent you should believe in the truth of the selected hypothesis), were de-emphasized for their perceived impossibility to account for all the relevant facts. Thus, while modern null hypothesis significance testing has adopted aspects of the decision-based nature of classical hypothesis testing, its roots as a method for statistical inference and making epistemic judgements based on probabilities about hypotheses hails from classical significance testing.

2.1.4 Students' Struggles and Criticism

To create a method that both specified a decision rule and made epistemic judgements, many researchers fused the decision error minimization approach with the significant difference approach to form null hypothesis significance testing (NHST; Lenhard, 2006). Perhaps unsurprisingly, this has led to much confusion among students,

statisticians, and textbook writers (Gigerenzer, 2004). The underlying logic of NHST has also been critiqued by researchers and practitioners along with the hypothetico-deductive reasoning it is based upon, being described as "bone-headedly misguided" (Rozeboom, 1997, p. 335). Forgotten in this fusion is the fact that Fisher as well as Neyman and Pearson vehemently disagreed with each other's approaches, with Fisher even suggesting a limited role for the hybrid NHST in statistical inference (Rao, 1992). However, by the late 20th century, NHST could be commonly found across introductory statistics textbooks (Nickerson, 2000).

Whether as part of NHST or truer to the early 20th century version of the significance test, thinking about null models is not trivial for students. Confusion and misconceptions have even been found in several textbooks and among statistics instructors and statisticians (e.g., Brewer, 1985; Falk & Greenbaum, 1986; Haller & Krauss, 2002; Mittag & Thompson, 2000). A review by Castro Sotos et al. (2007) of empirical research conducted between 1990 and 2006 identified five major struggles students had and common errors they made related to aspects of NHST concerning the significance testing approach:

(1) Misunderstanding the logic of the test, i.e., the conditional nature of the hypothetico-deductive approach, by incorrectly conditioning on the observed evidence;

(2) Difficulty specifying hypotheses and conflating hypotheses with decision rules;

(3) Misinterpreting *p*-values as the strength of an effect;

(4) Misunderstanding the inherent uncertainty in inference and viewing test results as deterministic; and

(5) Conflating statistical and practical significance.

Nickerson (2000) found many of the same beliefs among researchers and published papers in the psychological sciences. Additionally, Nickerson found some papers purporting a belief that failing to reject the null hypothesis is equivalent to proving it true, or that rejecting a null hypothesis proves a theory that predicted it would be false (both of which are theoretically erroneous).

Many of the recommendations Nickerson (2000) found being advocated for in the then current literature harken back to the core logic and procedures espoused by early 20th century statisticians such as Fisher and Neyman but forgotten through the decades. For example, using non-nil null hypotheses or providing specific alternative hypotheses were requirements in significance testing and hypothesis testing respectively as originally specified, but fell out of use in NHST. Further recommendations advocate distinguishing between the substantive contextual research question and the statistical hypothesis, re-emphasizing the role that a researcher's intuition plays in inference, just as it was emphasized by Fisher (1925) and Neyman and Pearson (1928).

Some critiques go further at striking at the underlying philosophy and recommend Bayesian approaches or emphasize abduction through likelihood-based inference (e.g., Rozeboom, 1997). Despite these disagreements about what to teach and how to teach it, only recently have calls been made to eliminate statistical procedures for the evaluation of hypotheses, instead focusing entirely on estimation (Cumming, 2014).

Nevertheless, significance testing and hypothesis testing continue to be a central part of statistical inference and a key learning goal in modern introductory level curricula (GAISE, 2016). Furthermore, the core purpose of significance testing, to probabilistically

reason about the relation between a candidate hypothesis and observed evidence, is fundamental to all statistical inference across all schools of thought and all interpretations of probability.

2.1.5 Simulation-Based Significance Testing

With the advent of modern computing in the late 20th century, statistics educators began calling for the use of simulation in the classroom to supplement the traditional mathematical aspects of statistical inference which students often found difficult to comprehend (e.g., Glencross, 1988). Null models historically took the form of probability models specified with parametric probability distributions (e.g., Z, T, $\chi 2$, F). After selecting the appropriate probability distribution as a null model, the significance level (i.e., the threshold beyond which observed evidence would, in relation to the null hypothesis, present a logical disjunction) could be identified by referring to regularly published tables or via manual calculations.

By the late 20th century, automatic hand-held calculators replaced probability tables (Moore, 1992), and some statistics educators saw in simulation-based methods an opportunity to reconsider the way students are taught the core ideas of statistical inference (e.g., Ernst, 2004). Simulation-based methods utilize simulators to generate probability distributions, rather than parametrically defined probability distributions. Compared to these parametric-based predecessors, simulation-based methods were thought to be more conceptually accessible for students (e.g., Budgett et al., 2013).

In a paper denouncing the then consensus introductory statistics curriculum as obfuscatory, costly, and fraudulent , Cobb (2007) articulated a set of core principles known as the three R's – "Randomize data production, Repeat by simulation to see what's typical,

and Reject any model that puts your data in its tail" (p. 13). These principles served as a basis for what is known as simulation-based inference (SBI) in introductory statistics curricula (Rossman & Chance, 2014).

Simulation-based inference generally refers to the use of a statistical model defined by a simulator to perform statistical inference (Cranmer et al., 2020). A simulator is any tool that can enact simulation, or in a statistical context, any computational algorithm that defines a population and assumes a data generating process to randomly generate multiple sets of sample data (Carsey & Harden, 2014). Simulators for statistical inference utilize resampling, a statistical procedure that reuses data from an observed sample in the service of statistical inference (Chernick, 2012).

While there are many types of resampling techniques such as bootstrapping, jackknifing, cross-validation, and randomization, introductory curricula overwhelmingly focus on bootstrapping and randomization (Brown, 2019). A bootstrap resampling procedure iteratively samples with replacement from the original dataset until the original sample size is reached. This process is used to mimic the process of random sampling from a population, with the original sample operating as an estimate of the population distribution. A randomization resampling procedure rearranges or regroups observations from the original dataset, a process used to mimic the process of random assignment into experimental groups.

Compared to its parametric-based predecessor, simulation-based inference generates an approximation of the null model via simulation rather than an approximation based on an algebraically derived theoretical probability distribution.

Seizing upon the pedagogical potential of simulation, statistics educators began creating ad-hoc simulation-based tools to develop students' understanding of various introductory level concepts in the late 20th century (e.g., delMas et al., 1999). Echoing these efforts, Cobb (2007) called on statistics educators to utilize simulation to free themselves and their students from the technical complexity and burden of expressing the sampling distribution mathematically, arguing that 21st century statistics instruction need not tether itself to the parametric-based methods once utilized out of necessity.

Before long, collections of activities became entire curricula, and simulation applications and software were either adopted or specifically designed for SBI curricula. The Rossman-Chance applets initially developed by Chance and Rossman (2006) were incorporated into the *Introduction to Statistical Investigations* curriculum (ISI; Tintle et al., 2020), the *Change Agents for Teaching and Learning Statistics* (CATALST) curriculum was developed around the TinkerPlots software (Konold & Miller, 2005; Zieffler et al., 2021), and StatKey was designed specifically for the *Statistics: UnLOCKing the Power of Data* curriculum (Lock5; Lock et al., 2021; Morgan et al., 2014).

These SBI software tools vary greatly in the degree to which users specify the characteristics of a simulator. For example, TinkerPlots requires users to construct models using a variety of resampling devices such as mixers and spinners (see Justice et al., 2018, for a detailed explanation of the user interface of TinkerPlots), whereas StatKey requires users to select one of several models identified by their functionality (e.g., Bootstrap Confidence Interval for a Single Mean, Randomization Hypothesis Test for a Difference in Proportions). Furthermore, each curriculum approaches simulation-based inference in a different way: ISI focuses on randomization before connecting simulation to parametric-

based methods (Tintle et al., 2011); Lock5 focuses on bootstrapping before introducing randomization and ultimately making the connection to parametric-based methods (Lock et al., 2021); and CATALST focuses on model creation through model eliciting activities (MEAs) before introducing randomization (and notably does not introduce students to parametric-based methods; Garfield et al., 2012; Justice et al., 2020). Despite their differences, all three curricula emphasize the role of simulators in statistical inference (and additionally all utilize active learning methods to promote student learning).

While simulation provides a different method of generating a null model than its parametric-based predecessor, simulation-based inference differs from Fisher's approach only in using a simulator as opposed to mathematical derivations when determining the expected frequencies with which different results for an experiment may occur. Cobb's second R, "Repeat by simulation to see what is typical" (Cobb, 2007, p. 13) satisfies the requirement of Fisher's "sampling distribution completely determined by the null hypothesis" (Fisher, 1935, p. 192). Cobb's three Rs were not a repudiation of Fisher, but rather a return to Fisher's core principles for inference and the importance of randomization. For example, Cobb's third R, "Reject any model that puts your data in its tail" (Cobb, 2007, p. 13) is based on the same thinking as rejecting a null hypothesis based on Fisher's "rational grounds for the disbelief it engenders" (Fisher, 1956, p. 43).

2.1.6 Summary of the Philosophy and Logic of Statistical Testing

What is statistical testing? A statistical test is a method of comparing the probabilistic consequences of a given hypothesis to observed evidence, to make inferences about the world.

What is the core logic of a statistical test? The key feature that makes a test a statistical test is the selection of a hypothesis that specifies a probability distribution for the possible outcomes of a study, which is called a null model. This exact criterion is what makes a hypothesis a null hypothesis, and what makes the test a statistical test.

Does the core logic of a statistical test change when statistical tests are conducted with SBI methods? No. The core difference between simulation-based approaches to significance testing and their parametric-based predecessors is that a simulator represents a data generating process under a null hypothesis, rather than a parametric probability distribution explicated through equations (Cobb, 2007; Fisher, 1935). Thus, in SBI, the key feature that allows a null hypothesis to be tested is its simulator that specifies an underlying null model. With increased training efforts and a nascent evidence basis, simulation-based methods appear to be ascendent as a pedagogical tool for introductory level statistics, placing increased importance on students' thinking about null models and null model simulators.

However, the classical approach to significance testing utilizing null hypotheses, null models, and *p*-values have long drawn criticism from practitioners and theorists alike (Cohen, 1994). Debates over their utility and appropriateness have occurred almost continually since their formalization in the early 20th century (e.g., Berkson, 1938; Gigerenzer, 1993; Hogben, 1957; Morrison & Henkel, 1970; Nickerson, 2000; Wasserstein et al., 2019). Beyond students' difficulty in learning the method, much of the critique of the classical approach to significance testing has been centered around either its process of drawing conclusions or its underlying logic (e.g., Gigerenzer, 2004; Rozeboom, 1997). Despite such controversies, significance testing continues to be a central part of statistical

inference and a key learning goal in modern introductory level curricula, including simulation-based curricula (GAISE, 2016).

2.2 Students' Difficulties and Understanding with Simulation-Based Inference

The late 20th century and early 21st century have seen the emergence of simulation-based methods and software tools to teach statistical inference, replacing the parametric-based methods and probability tables of the early 20th century (Rossman & Chance, 2014), all in an attempt to alleviate students' difficulties by placing the logic of statistical inference at the core of instruction (Cobb, 2007). These new tools have led to the development of new SBI curricula which have, especially since the early 2010s, begun to be rigorously evaluated. Preliminary evidence suggests that these new SBI curricula may lead to marginal improvements in students' understanding (Brown, 2019). However, despite some observed changes in the way students reason about statistical inference with simulation (Case, 2016), some evidence suggests that many students still struggle to apply, explain, and justify inferential procedures with these SBI methods (e.g., Noll & Kirin, 2017).

With increased training efforts and a nascent evidence basis, simulation-based methods appear to be ascendent as a pedagogical tool for introductory level statistics. The core difference between these simulation-based approaches to significance testing and their parametric-based predecessors is a simulator representing a data generating process under a null hypothesis (Cobb, 2007; Fisher, 1935). The key feature that allows a null hypothesis to be tested is its simulator that specifies an underlying null model. As more simulation-based curricula and software tools are developed, and as calls for the reform of statistics instruction explicitly recommend the elimination of significance testing, there is a need for research examining students' understanding of significance tests in simulation-based

curricula and of null models in SBI. In particular, (1) to what extent do students' difficulties, documented within parametric-based statistics curricula, persist even in simulation-based curricula, (2) what is students' understanding of null model simulators, which are core to the logic of the simulation-based statistical test, and (3) what if any unique aspects to students' thinking about significance testing emerge within simulation-based pedagogies.

The first studies evaluating curricula primarily teaching simulation-based inference occurred in the early 2010's. Since then, three general types of evidence have been the focus when examining students' understanding and thinking: (1) students' responses to forced-choice assessment items (e.g., Tintle et al., 2011), (2) students' responses to constructed-response assessment items and other written assignments (e.g., Frischemeier & Biehler, 2013), and (3) observations and interviews of students when conducting simulation-based inference tasks (e.g., Noll et al., 2018a). Together, this diverse body of evidence suggests that curricula primarily teaching simulation-based inference may lead to higher gains in students' understanding of significance tests. However, students may not understand null models and simulators as well as they understand how to draw conclusions from a significance test. Furthermore, students' thinking about null model simulators appears more complex than current theories and conjectures account for. The next section summarizes relevant evidence assessing students' understanding before discussing evidence related to their reasoning and thinking.

2.2.1 Students' Understanding of Simulation-Based Significance Testing

Preliminary results from comparative studies have generally found that, on average, students' gains in conceptual understanding are higher with introductory level simulation-

based curricula than traditional parametric-based curricula. A review by Brown (2019) identified 13 multi-classroom studies comparing student learning outcomes between classes, curricula, or in comparison to results from a previously published study. All but one of these studies had group sample sizes of at least 100, with two large scale studies including over 10,000 total participants (Chance et al., 2018; VanderStoep et al., 2018). All studies used the Comprehensive Assessment of Outcomes in Statistics (CAOS; delMas et al., 2007) or a modified version of it to assess student learning outcomes. In general, Brown (2019) found that students in simulation-based inference groups performed no worse than traditional inference groups in terms of their total scores on these assessments. Notably, VanderStoep et al. (2018) found that when stratifying students into three groups by their pretest scores, gains in overall understanding were higher for students in the ISI curriculum for students in the low and middle pretest score groups, and Chance et al. (2016) found that students' gains in understanding were comparable for students with instructors both new to simulation-based curricula and instructors more experienced with SBI.

The use of simulation also appears to shape the way in which students understand statistical inference. In one of the only comparative qualitative studies of high school students' understanding of statistical inference in both traditional methods and simulation-based methods, Case (2016) found that students perceived traditional methods to be an easier procedure to enact, exemplified by one student's comment that "if you know when to do the test and you know how to do the test, you don't really have to understand what you're doing" (p. 93). Case also noted differences in students' interactions with the various tools of traditional inference, predominantly the graphing calculator, and the tools of simulation-based inference, either physical manipulables or software, and suggested that

these differences shaped how students learned and what they understood about statistical inference. Noll and Kirin (2016) similarly noted that utilizing the TinkerPlots software framed students' thinking when approaching statistical inference tasks. This section next explores how simulation-based inference affects students' understanding of null models with significance testing.

2.2.1.1 Students' Understanding of Null Models

In seven of the studies identified by Brown (2019) that evaluated the ISI curriculum, students' scores on Tests of Significance items were consistently higher for students in the ISI simulation-based curriculum than in traditional parametric-based curricula (see Table 1). Furthermore, students' scores from those in the ISI curriculum were generally higher on average for Tests of Significance items than the Confidence Interval items and the Sampling Variability items, which focused on general ideas about sampling variability such as the law of large numbers.

Hildreth et al. (2018) found a similar pattern when comparing students' scores from sections utilizing the CATALST curriculum and the Lock5 curriculum to two traditional parametric-based curricula – the average posttest score on three items from CAOS measuring understanding of *p*-values for 1584 students in traditional curricula was 62.8% (see Appendix A for more about these CAOS items), while the average posttest score on the same items was 82.1% for 770 students in the CATALST curriculum and 83.8% for 758 students in the Lock5 curriculum. Similar results in overall comparisons between curricula were found by Garfield et al. (2012) when comparing the CATALST curriculum to a traditional parametric-based curriculum. Garfield et al. utilized the Goals Outcomes Associates with Learning Statistics assessment (GOALS), based on sixteen items from

CAOS but with an additional seven items explicitly focusing on the use of simulation methods to draw inferences. Average student scores were higher for students in the CATALST curriculum, and were also, on average, higher for the seven simulation-based inference items compared to three items related to confidence intervals and four items related to sampling variability.

It is important to note that the assessment used by all these studies was either CAOS or a derivative of it. All six CAOS items in the Tests of Significance topic and all nine items in the corresponding ISI assessment topic only concern the correct interpretation of *p*-values and the decisions to be made based on this result, corresponding with Cobb's third R, Reject (see Appendix A). While there are items on both assessments that assess students' understanding of the purpose of randomization, these items do not explicitly relate to the relationship between random processes and a null hypothesis, and thus do not relate to null model simulators or null models in any direct manner. Similarly, there are no items on either assessment that address Cobb's second R, Repeat by simulation.

Studies explicitly examining students' understanding of simulation in significance testing appear to identify gaps between students' understanding of the interpretation of the results of a significance test, their understanding of study design characteristics' relation to appropriate conclusions, and their understanding of null models and the role simulators play in simulation-based significance tests. A later version of the GOALS instrument, GOALS-4, was utilized by Sabbag et al. (2015) and consisted of 20 total items, five of which assessed students' reasoning about *p*-values (based on items from CAOS), and two of which assessed students' understanding of null models (see Appendix B). Students'

scores were on average lower for the two items assessing understanding of null models than the for the five *p*-value items.

Frischemeier and Biehler (2013) found similar evidence suggesting that students may not understand null models as clearly as they understand interpreting *p*-values. They provided their students, pre-service mathematics teachers, with a randomization test plan (Table 2) consisting of six steps to help support the structural aspects of their thinking along with an example solution to the Extra Sensory Perception (ESP) task (Rossman et al., 2001). After completing the ESP task, Frischemeier and Biehler (2013) studied students' use of TinkerPlots when performing randomization tests on the Muffins task (Biehler et al., 2003). They analyzed submitted written work from 11 student pairs at the end of the course and compared these responses to expected correct solutions, rating each step of the plan as successfully completed or not. They found that 10 out of 11 teams correctly created a null model using TinkerPlots, even though only 8 out of 11 teams correctly formulated a null hypothesis. Students were similarly successful in specifying the test statistic (10 out of 11) but struggled with successfully calculating a *p*-value and drawing conclusions from the test based on the null model they specified (5 out of 11 in each step).

Taken together this evidence suggests that understanding how to draw conclusions from significance tests may not imply an understanding of the central role null models play in significance tests nor the essential role of null model simulators in simulation-based tests. Furthermore, it suggests that understanding how to interpret *p*-values may not imply an understanding of the relationship between *p*-values and their underlying null models. However, such an interpretation of this evidence is tenuous at best – inferences based on students' responses to different GOALS-4 items depend on the marginal reliability and

distinctness of these items when measuring differences in students' understanding, and inferences based on the frequency of correct responses to written tasks depend on the quality of the rubric distinguishing between correct and incorrect responses. Nevertheless, differences in average correct responses on each item in GOALS-4 and the varying proportion of correct responses according to Frischemeier and Biehler's (2013) randomization test plan highlight that there may be gaps in students' understanding. Yet, neither study provides evidence of students' reasoning and thinking that may identify which of the parts of conducting a significance test students may have struggles with, nor how they conceptualize the task of conducting a significance test and if students' thinking at all differs from researchers' expectations.

2.2.1.2 Students' Creation of Null Model Simulators

Some students' difficulties in conducting simulation-based significance tests may be explained by students' struggles in utilizing simulation software such as TinkerPlots to model the exact characteristics of a null hypothesis and its underlying null model. To further explore how students understand null models, both in terms of its function statistically as well as how to utilize the TinkerPlots software to successfully create a null model simulator, Biehler et al. (2015) examined submitted written work from 18 pairs of pre-service mathematics teachers, who were recruited two months after they completed an introductory statistics course.

These students were then given the Verdienststrukturerhebung [Structure of Earnings Survey] (VSE) task based on data collected by the German Statistisches Bundesamt [Statistics Bureau] (Figure 1). Students were asked to fill out a blank randomization test scheme, a slightly modified version of Frischemeier and Biehler's

(2013) test plan, and their answers were rated as 'successful' if they adhered to expected solutions. Approximately 89% of participants (16 of 18 pairs) correctly specified a null hypothesis (step 2). However, Biehler et al. (2015) found that most participants did not provide a clear description of how the null hypothesis would be translated into a null model with TinkerPlots, and due to large variation in students' responses did not rate them as successful or not. Furthermore, they found that participants struggled with the initial creation of the null model in TinkerPlots – only 56% of participants (10 of 18 pairs) correctly populated the sampler, 72% (13 of 18 pairs) correctly set the number of repetitions, and 50% (9 of 18 pairs) correctly specified sampling without replacement. Yet, this may not be due to a lack of understanding about a null model, as Maxara and Biehler (2007) found evidence that students struggle translating probabilistic models into TinkerPlots outside the context of a null model and significance testing.

Given students' ability to correctly specify a null hypothesis in both Biehler et al.'s (2015) study and Frischemeier and Biehler's (2013) study, and their difficulty correctly specifying a null model simulator with TinkerPlots as found by Biehler et al. (2015), two potential explanations emerge – students struggle to utilize TinkerPlots to realize their conceptually well-defined null models, or perhaps they struggle to understand the relationship between the null hypothesis and its underlying null model. Furthermore, as both Frischemeier and Biehler (2013) and Biehler et al. (2015) provided the randomization test plan to students as part of their instruction, it may also be that the plan is difficult for students to learn as part of a learning trajectory, or that students' thinking about significance tests and their internal schema for conducting such tests are not isomorphic to this test plan.

Explicitly building off the work by Biehler et al. (2015), Noll and Kirin (2017) sought to explore why students struggled with the initial creation of the null model simulator in TinkerPlots. They observed students while solving the Dolphin Therapy task (Figure 2), and focused their analysis on the three TinkerPlots steps associated with populating a mixer, setting the number of repetitions, and specifying replacement. Noll and Kirin found that populating the mixer was directly linked to students' interpretations of the null hypothesis, and students did not specifically deliberate this point outside of discussions about the null hypothesis. Students were also intuitively able to specify the correct number of repetitions based on the total sample size.

However, students appeared to struggle with determining whether the device should be set to 'with replacement' or 'without replacement'. Two groups of students who correctly chose 'without replacement' simply compared this task to a previous activity, without explicitly acknowledging that this selection allowed the TinkerPlots device to mimic the random allocation process. The two groups who incorrectly chose 'with replacement' did so for different reasons. One group desired to maintain the same chance of improving for each individual in the device. The other group hoped to model a bootstrap resampling process to facilitate generalizations of their results, despite bootstrap resampling being incongruous with the original study design of the problem.

Noll and Kirin (2017) also noted that students struggled to operationalize the null hypothesis's statement of 'no difference' at the group level and instead modeled an equal chance of improving or not improving for each individual. This struggle was also noted by Biehler et al. (2015), despite students being able to correctly specify a null hypothesis utilizing mathematical symbols. However, these struggles may simply be due to the

idiosyncratic features of specifying 'replacement' in TinkerPlots, which may be unintuitive for students as they learn to conduct significance tests and conceptualize null models. Nevertheless, together these two studies suggest students may struggle to utilize TinkerPlots to transcribe null hypotheses into null model simulators. Furthermore, the correct specification of a null hypothesis with mathematical symbols does not appear to imply that students can correctly operationalize the null hypothesis as a null model.

Despite these errors in correctly specifying the characteristics of a null model simulator and justifying these choices, Noll and Kirin (2016) found evidence that students are generally able to utilize TinkerPlots to create their intended null model simulators. To provide a more detailed account of how students relate aspects of statistical problems to the TinkerPlots models they construct, Noll and Kirin (2016) used an inductive coding method to analyze students' written work on a significance testing task. They examined undergraduate non-statistics major students' construction of models in TinkerPlots by evaluating their answers to a question from the Models of Statistical Thinking assessment (MOST; Garfield et al., 2012), called the Facebook Task (Figure 3). Despite large variability in the types of models students created, only seven of 33 students justified their design choices in a manner contradictory to the features of their constructed device, while 23 students created models consistent with their justification and explanations (although three of these students created incorrect models). This evidence appears to imply that within the context of the Facebook task, students were able to use TinkerPlots to realize their desired models.

Building on the findings of Noll and Kirin (2016), Noll et al. (2016) found evidence that students generally understand the role that a null model simulator plays in simulation-

based inference. Analyzing the Facebook task again, along with the Music Note task (Figure 4), Noll et al. (2016) examined students' explanations in terms of how they related to four conjectured phases of inferential reasoning. Noll et al. (2016) hypothesized that students' thinking occurred in four phases in which students: (a) appropriately construct a TinkerPlots model corresponding to the null hypothesis; (b) use the model to generate a single trial and suitably represent its outcome; (c) generate multiple trials to create a distribution (i.e., the null model); and (d) utilize the results from all three previous phases to draw conclusions. Noll et al. found that even when students struggled to construct an appropriate TinkerPlots model corresponding to the null hypothesis, they were able to enact simulations and reason about the null model correctly. Two common errors were incorrectly assuming that 'by chance' implies a probability of .5 and designing a null model simulator based on the observed results rather than the specifications of a null hypothesis.

It should be noted that the Facebook task and Music Note task are both theoretically isomorphic and have a substantially different study design than the Dolphin Therapy task, which requires different settings in TinkerPlots. Specifically, students do not have to specify 'without replacement' for the Facebook task, a setting which proved difficult for students studied by Noll and Kirin (2017). Therefore, students' ability to utilize TinkerPlots to realize their intended models may be limited both by gaps in their understanding of the full TinkerPlots functionality as well as gaps in understanding null hypotheses across various contexts and statistical content. Similarly, their understanding of the role of simulation and their ability to enact simulation may only be limited to situations in which the resampling method required is bootstrap resampling, as is the case in the Facebook task and Music Note task. For example, one group of students studied by Noll and Kirin (2017)

incorrectly attempted to utilize bootstrap resampling in the Dolphin Therapy task, despite the problem requiring randomization resampling. Nevertheless, it appears that, at least in some scenarios and tasks, students are able to successfully create null model simulators with TinkerPlots.

2.2.1.3 Summary of Students' Understanding of Null Models and Null Model Simulators

Preliminary evidence from evaluations of students in simulation-based curricula suggest that simulation-based inference may improve students' understanding of significance testing compared to the traditional parametric-based inference, particularly in terms of interpreting *p*-values (e.g., Hildreth et al., 2018; VanderStoep et al., 2018). Yet, an understanding of the conclusions that can be drawn from a significance test may not imply that students also understand the role that null models play in significance testing (Frischemeier & Biehler, 2013; Sabbag et al., 2015). While students struggle to create null model simulators in TinkerPlots (Biehler et al., 2015; Noll & Kirin, 2017), for some problems they seem generally able to use TinkerPlots to represent their intended models (Noll & Kirin, 2016). Furthermore, within some problem contexts, students seem to generally understand the role of simulation with regard to the null model simulator and significance testing (Noll et al., 2016). Therefore, students' struggles with simulation-based significance testing may be due to:

(1) idiosyncrasies of particular problem contexts and types of statistical study designs,

(2) difficulties extracting a null hypothesis from context,

(3) converting the null hypothesis into a specific null model simulator, or

(4) ensuring that the null model simulator contains complete information to facilitate the creation of a null model in TinkerPlots.

Although converting a null hypothesis into a null model is not a problem unique to simulation-based significance testing, simulation appears to inform the way students approach statistical inference (Case, 2016). There is evidence that when taught with curricula that utilize TinkerPlots, students approach significance testing tasks with TinkerPlots models in mind (Garfield et al., 2012; Noll & Kirin, 2016). Thus, simulation-based curricula, especially those utilizing TinkerPlots or similar software that make explicit students' representations of the null model in the form of a null model simulator, may be able to provide researchers an insight into students' thinking and their processing of contextual and statistical information in significance testing tasks. However, the role these software tools play in shaping students' understanding of significance testing remains largely unexplored.

2.2.2 Students' Thinking about Null Model Simulators

Despite difficulties and errors that students make when conducting simulation-based significance tests (e.g., Biehler et al., 2015), some evidence suggests that students are, under certain circumstances, generally able to utilize TinkerPlots to create intended models and understand the role null model simulators play in SBI (Noll & Kirin, 2016; Noll et al., 2016). Studies by Noll and Kirin (2016) and Noll et al. (2018b) found large variability in the types of TinkerPlots models students created when solving the Facebook task (see Figure 3) and the NFL task (Figure 5) respectively. One possible explanation for students' errors is difficulty in extracting a null hypothesis from a problem context and the

complete characteristics manifested in the null model simulator that the null hypothesis specifies.

Beyond difficulties in specifying TinkerPlots sampler characteristics such as draw, repeat, and replacement, approximately half of Noll and Kirin's (2016) students working on the Facebook task created a single device that focused only on the day of the break-up, while the other half created a linked device that separately accounted for individual couples that broke up and day of the break-up. Similarly, in completing the NFL task, approximately half of Noll et al.'s (2018b) students created a single device focusing on either the winner of the coin flip or the winner of the game exclusively, while the other half created a linked device that separately accounted for both. These tasks are mathematically isomorphic – 'Given a breakup has occurred, what is the chance it occurred on Monday?' is equivalent to asking, 'Given a team has won the coin flip, what is the chance it will win the game?'. Yet, students creating linked-devices in the Facebook task had more errors and difficulties specifying the device, while students creating single devices in the NFL task had more errors and difficulties specifying the device. In both cases, students creating linked devices experienced difficulty analyzing and summarizing the results from their samplers and struggled to account for the conditional nature of the task.

Two studies by Noll et al. (2018a, 2023) suggest that an inherent human predisposition for narrative sensemaking may explain how students construct null model simulators as well as how they interpret them. In general, narrative thinking and reasoning processes can help students organize information into a coherent structure (e.g., Clark & Rossiter, 2008). Furthermore, statistical models may be inherently narrative, as they "bring

forth important aspects of a problem, contain an underlying statistical structure of a process, and are purposeful (used to make sense of a problem)" (Noll et al., 2018a, p. 1269).

In examining video-recordings of students completing the Music Note task (see Figure 4), Noll et al. (2018a) noticed that students appeared to focus on narrative characteristics of the problem context when constructing TinkerPlots models. For example, many students constructed TinkerPlots models based on the temporal sequence of the problem context, ensuring that their models accurately reflected that "the teacher plays the note … and then the student guesses it. And so, it's not at the same time" (p. 1274). This strong link between the TinkerPlots model (and the null hypothesis it is meant to represent) and the original problem context could also produce a narrative tension until subsequent contextual details were added by the students. One student highlighted this tension by stating "My only problem with this, is that it doesn't really put into play what the student really knows about the music" (p. 1274) to which a group mate responded that under the instructions of the problem "the student doesn't really know anything about music, so it's totally random" (p. 1274-1275). Students also valued TinkerPlots models for their communicative power, preferring models that they perceived accurately told the story of the problem task. This evidence suggests that students' creation of null model simulators is dependent on a process that integrates both a specific null hypothesis and the story structure of the problem context, and successfully resolves tensions between the two.

Even when students are able to successfully create a null model simulator, they may not understand the random and repeatable process it represents. A study of in-service statistics teachers by Justice et al. (2018) utilized structured interviews to examine teachers' understanding of null model simulators as a data generating process (DGP). A DGP

approach to null model simulators exemplifies Konold et al.'s (2007) theory of understanding distributions through modeling and the core principle of randomization as a data production process (Cobb, 2007; Fisher, 1935). The teachers in the study explicitly focused on creating null model simulators that replicated the manner in which the original data was produced, emphasizing some elements such as the temporal sequence of the study that do not affect the simulation results. In doing so, Justice et al.'s participants unanimously viewed the null model simulator's role as facilitating comparisons between the hypothesis and the evidence that could lead to an inference or conclusion. While this comparison is fundamental to the task of a significance test, it emphasizes the null model simulator's by-product (i.e., the null model, and ultimately, the *p*-value) rather than the random data generating process that it represents, and which is important to think about when conceptualizing the null model simulator's core purpose and role in significance testing. However, while these teachers teach the CATALST curriculum, they were not trained with it as students, and their understanding of null model simulators in relation to the null model may be a residual effect of their parametric-based statistical training.

2.2.3 Summary of Students' Understanding and Thinking about Simulation-Based Significance Testing and Null Model Simulators

Simulation-based inference has captivated statistics educators through several hypothesized benefits (e.g., Cobb, 2007). Simulation-based curricula such as ISI, CATALST, and Lock5, may lead to higher gains in students' understanding related to drawing conclusions from significance tests and interpreting *p*-values, a task that has historically befuddled students as well as some instructors (e.g., Nickerson, 2000). However, simulation-based approaches also entail unique aspects to conducting

significance testing, namely in the specification and utilization of a null model simulator in simulation software such as TinkerPlots.

While students are, in certain circumstances, generally able to understand the roll a TinkerPlots model plays in significance tests (Noll et al., 2016) and are sufficiently fluent with TinkerPlots to specify their intended models (Noll & Kirin, 2016), students struggle to operationalize the null hypothesis as a null model simulator in TinkerPlots (e.g., Biehler et al., 2015; Noll & Kirin, 2017). Students struggle to resolve tensions between statistical hypotheses and characteristics of the problem context such as the temporal sequence of events or their personal beliefs about what should or should not affect the results of a study (Justice et al., 2018; Noll et al., 2018a). This struggle to integrate the null hypothesis with the problem context leads to a wide variety of operationalizations of the null model simulator (Noll & Kirin, 2016; Noll et al., 2018b), and a strong preference for simulators that communicate 'the story' of the original study and strictly adhere to its design (Justice et al., 2018; Noll et al., 2018a).

Furthermore, students may view the purpose of a null model simulator only in terms of its product, the null model and *p*-value attained once simulation is enacted, and not in terms of the random and repeatable data generating process as specified by the null hypothesis that the null model simulator represents (Justice et al., 2018). Therefore, while simulation-based methods do appear to improve some aspects of students' understanding of and thinking about significance testing, there is a need for further research aimed at developing students' ability to transcribe a null hypothesis into a null model simulator, emphasizing the intermediary importance of the exactness of the null model, and its purpose in significance testing.

2.3 Summary of Current Research on Students' Thinking about Null Models

One of the key tools of science is the evaluation of hypotheses through experimentation and testing. Statistical inference explicitly aims to utilize probability when thinking and reasoning about the strength of inferences about such hypotheses. While there are many schools of thought on how to approach statistical tests, one of the most common methods found throughout textbooks and 20th century practice is a significance test, or its derivative method, null hypothesis significance testing (NHST). To conduct a significance test, an individual first specifies an exact probability distribution for possible outcomes of an experiment based on a candidate hypothesis, which is known as a null model under and is based on a null hypothesis. Then, observed evidence is compared to this null model, and when the observed evidence differs from the expectations based on the null model, the candidate hypothesis is considered (probabilistically) nullified. This thinking and reasoning follows the hypothetico-deductive approach to confirmation, with the added element of the specification and utilization of probability.

Beyond methodological and philosophical critiques of this method, students have historically struggled to conduct this significance testing procedure. However, new curricula based on simulation appear to lead to larger gains in students' understanding about significance testing than their parametric-based predecessors. These simulation-based methods present their own challenges for students, with students often struggling to transcribe a null hypothesis into a null model simulator that subsequently can generate the null model. As these simulation-based methods appear to frame students' understanding and approach to significance testing, new research is required to understand how students

grapple with and think about this unique intersection of simulation software and statistical hypotheses.

Biehler et al. (2015) proposed a framework explicating the thinking required to complete simulation-based significance testing (Figure 6). The framework describes three levels successively dependent on each other – a statistics level dependent on the problem context and a software level dependent on the statistics level. It is in completing the transition from the problem context to the statistics level that students have the greatest difficulty – extracting a statistical problem from a given task, determining an appropriate null hypothesis, and explicating a null model simulator that specifies all characteristics to be transcribed from the null hypothesis to generate a null model (e.g., Noll & Kirin, 2016; Noll et al., 2016). However, it is unclear whether students' difficulties creating null model simulators are due to a lack of conceptual understanding of null hypotheses and null models or difficulty in processing problem tasks as presented to them.

Noll et al. (2018a, 2023) found evidence that students were sensitive to the narrative structure of a task. When reading a text, individuals construct a referential situation model of what the text is about utilizing working memory and based on the text's characteristics (Graesser et al., 1994). Working memory is the cognitive mechanism that facilitates the storing and processing of information (Baddeley, 2003). Working memory both processes information and facilitates long term storage of information, and thus there is a trade-off in the efficiency of working memory to simultaneously achieve both (McCutchen, 2000). For example, increased organizational structure of information can support improvement in text recall (Meyer & Freedle, 1984). Students' difficulties resolving tensions between the

null hypothesis and the problem context may thus be a function of the text characteristics of the problem task provided to them.

It is worthwhile to reiterate that many students are novices with regards to formally reasoning about probability and statistics. When considering the Facebook task (see Figure 3), the Music Note task (see Figure 4), and the NFL task (see Figure 5), experts might readily see the same mechanism and statistical structure in each task (e.g., a simple urn model). This translates to a single TinkerPlots device with two outcomes, the event of interest and its complement.

However, each of these tasks resulted in students producing a wide variety of TinkerPlots models and presented unique challenges for students in transcribing the null hypothesis into a null model simulator. While structurally isomorphic in terms of their statistical content and underlying probability models, the story or text structure of each task as presented to students was notably different. For example, the Music Note task separately describes how the music teacher plays a note at random before introducing how the student provides an answer or guesses the note played, while the NFL task does not separately discuss the coin flip procedure and the act of winning or losing the game (Noll & Kirin, 2016; Noll et al., 2018b). This variation in text characteristics may explain variation in the formulation of a null model and a null model simulator in students' responses.

Furthermore, students appear to prefer models with communicative power in relation to the problem task. Noll et al. (2018a) documented one student's view that "[the single device model] is more efficient if you're just trying to get the distribution, but if you want to like, tell the story of what happens, this [linked device model] or the two spinners more accurate displays what's actually happening" (p. 1277).

This preference for communicative models may also provide an opportunity to facilitate the development of students' understanding of null model simulators. By presenting significance testing tasks with text characteristics that embed statistical information essential to the exact specification of a null model and null model simulator as a natural part of the story of the task, students may be able to better abstract this information, facilitating transcription of the null hypothesis into a null model simulator. This may also present opportunities to scaffold students' understanding of null models by successively omitting statistically irrelevant information in problem tasks linked to the gradual suppression of information in a TinkerPlots model (i.e., collapsing a linked device to a single device, or condensing a mixer with all possible outcomes in the sample space to an urn model with only two outcomes).

Another possible explanation for students' struggles through the lens of Biehler et al.'s (2015) framework is that students appear to view the purpose of a null model simulator only in terms of its product (i.e., the results of the simulation), and not for its role as the representation of the null hypothesis nor the random and repeatable data generating process underlying the null model (e.g., Justice et al., 2018). While students are generally able to think and reason from the results of the simulation to statistical inferences and conclusions about the problem context, they struggle to reason from a real problem to a statistical problem and to the null model simulator (e.g., Noll et al., 2016). Biehler et al. (2015) postulated that some students may have an internal schema for randomization test procedures using TinkerPlots, but lack a conceptual understanding linking a null model simulator, in terms of both its represented process and its enacted product, with statistical inference and the logic of significance testing. While this theoretical framework specifies

a ‘statistical’ level, this may not be a distinct level in students’ minds, which may also explain their conflation of process and product when considering the purpose of a null model simulator.

2.3.1 Limitations of Current Research

One glaring limitation of the body of research examining in detail students’ thinking about significance testing in simulation-based curricula through the early 2020s is that nearly all of the studies utilized the TinkerPlots software, and have only been conducted in one of two curricula, either the CATALST curriculum or the curriculum developed by Frischemeier and Biehler (2013) and Biehler et al. (2015). Studies evaluating the ISI and Lock5 curricula primarily do so through the use of either CAOS or CAOS-based assessments which thus far have not included items specifically addressing students’ thinking about null models and null model simulators. Furthermore, the studies by Frischemeier and Biehler (2013) and Biehler et al. (2015) recruited pre-service mathematics teachers, while the studies by Noll and Kirin (2016, 2017) and Noll et al. (2016, 2018a, 2018b, 2023) recruited mostly liberal art majors, many of whom identified as poor math students (Noll & Kirin, 2016). Therefore, it is nearly impossible to disentangle curricular effects, individual and group differences, and the differences in research methods when making inferences about students’ understanding of and thinking about significance tests based on this body of empirical evidence. Nevertheless, these studies provide an important first glimpse at students’ understanding and thinking, and document for the empirical record students’ struggles in these simulation-based approaches to significance testing.

Perhaps it is by coincidence that most of the studies documenting students' performance on assessment items related to significance tests were evaluations of the ISI curriculum, while most of the studies documenting students' thinking about significance tests through observation or written work utilized the TinkerPlots software and the CATALST curriculum. While Hildreth et al. (2018) found that students' understanding of statistical inference was comparable between the CATALST, ISI, and Lock5 curricula, TinkerPlots requires its users to create null model simulators from scratch, as opposed to the Rossman-Chance applets or StatKey which provide pre-constructed null model simulators or require only partial specification of its characteristics. Although these software tools do not require students to explicitly construct null models, it does not mean they do not build an understanding of some or all aspects of the null model, as they still see and interact with the product of the enacted simulation, the null model.

While TinkerPlots provides a useful mechanism for researchers to observe students' creation of null model simulators and thus also of null models, the current body of research leaves open the question as to what learning benefits such explication has for students, and to what extent current research findings are simply the result of TinkerPlots's idiosyncrasies. As more simulation-based curricula emerge (e.g., Çetinkaya-Rundel & Hardin, 2021), and other simulation software are developed, it is important to verify that current research findings are not unique coincidences dependent on the specific curriculum or software utilized in previous studies, and if there are differences in students' understanding and thinking about significance tests across curricula and software tools, and to which curricular and software specifications these differences may be related.

Despite the propagation of simulation-based methods, relatively little is known about their effects on students' reasoning and thinking about null models. Many commonly used assessments of students' understanding do not explicitly include items that address students' understanding of null models and null model simulators. While such items are a part of the other assessments (e.g., GOALS, MOST, Garfield et al., 2012; Introductory Statistics Understanding and Discernment Outcomes assessment, I-STUDIO, Beckman, 2015), students' results utilizing these assessments have not been reported at the item level. Secondary data analyses can shed further light on the distinction between students' understanding of interpreting results from significance tests and their understanding of the role null models and simulation play in significance tests. Similarly, several assessments include items concerning study design and random processes such as random sampling and random allocation, but not explicitly in relation to null hypotheses (e.g., CAOS, delMas et al., 2007; Inferences from Design Assessment, IDEA, Fry, 2017). Including items that correspond to all three R's of Cobb's framework for simulation-based significance testing (i.e., 'Randomize', 'Repeat', 'Reject') can help shed further light on students' understanding and thinking about null model simulators and null models, which are core to the logic of significance testing, and thus the main emphasis of simulation-based pedagogies.

A focus on the relationship between study design and null models, meant to foster students' understanding of the data generating process that the null model simulator represents, also provides a unique opportunity to incorporate null models in Model Eliciting Activities (MEA; e.g., Garfield et al., 2012) which are already a part of some simulation-based curricula. While MEAs typically ask students to build a model based on

real-world patterns, a Null Model Eliciting Activity might instead focus on predicting real-world patterns based on a model. These Null Model Eliciting Activities may aid in the development of students' understanding of and thinking about the variability specified in null hypotheses and ultimately the transcription of a null hypothesis into a null model simulator.

Previous research on students' thinking about null model simulators also invites the use of several promising frameworks for future research. As noted by Noll et al. (2018a), humans use narratives and stories to help make sense of their experiences and organize their knowledge (Clark, 2010; Schank, 2000). Future research can investigate the role that narratives may play in how students create, make sense of, and understand null models in simulation-based software. Additionally, research can explore the potential of instructing students using statistical narratives and its effects on students' comprehension, processing, and recollection.

2.3.2 Problem Statement

Despite the propagation of simulation-based methods, relatively little is known about their effects on students' reasoning and thinking about null models, core to the logic of statistical testing. Furthermore, current studies have utilized a limited set of research methods and designs, and have also been used on a limited set of students. What is graduate students' thinking in statistical tests? Graduate students are a population that have historically struggled to learn statistical testing (Nickerson, 2000), and for whom there is a clear need to learn statistical testing (e.g., APA, 2017). What might a distal measure of students' thinking reveal (beyond the two month delay of Biehler et al., 2015)? How much of what students are taught do they remember? This is not a question of transfer, such as

that asked by Beckman (2015). Instead, this is a question of memory and lasting impact that may extend into graduate students' careers as scholars and researchers.

After an extended delay subsequent to the completion of an introductory statistics course using an SBI curriculum, what is graduate students' thinking in statistical tests?

2.4 Review of Theories and Methods for Researching Students' Statistical Thinking

To support the selection of research methods to investigate the aforementioned question, this section briefly reviews the background theories and evidence relevant to the objects of study (i.e., statistical tests and their null models), as well as the orienting framework for analysis and a brief review of methods utilized in similar studies.

2.4.1 Orienting Framework

To explicate and analyze students' thinking about significance testing, this study builds off of the work of Noll et al. (2018a, 2023) by utilizing statistical narrative as an orienting framework. Humans use narratives to help make sense of their experiences and use narration as a sense-making medium (Clark, 2010). Narratives are so pervasive to human cognition and the human experience that some have even dubbed our species homo narrans (Fisher, 1984; Rowe et al., 2007).

At its most fundamental, narrative is a way to help organize knowledge (Schank, 2000). Narratives organize a unique sequence of events or happenings, which are only given meaning through their place in an overall configuration, called a plot or fabula (Bruner, 1990). A plot is the basic scheme of events, which can be either thematic, structural, or a combination of both (Jahn et al., 2010). Narratives are more than just an amalgamation of episodes, instead communicating both plot and events together as one meaningful aggregate totality (Ricoeur, 2005). Narratives thus must have an overarching

conceptual scheme providing contextual meaning to individual events (i.e., a plot), and must draw together multiple events, happenings, or actions that are thematically unified to achieve a particular goal (Polkinghorne, 1995).

Statistical narratives are the stories we tell as part of the process of statistical thinking (Noll et al., 2018a). Statistical narratives integrate events from a problem context with statistical entities and software functionality. While specific research on students' statistical narratives is nascent, statisticians have long argued that statistical thinking resembles features of narrative reasoning and thinking. Cobb and Moore (1997) discuss statistics in terms of a dialog between statistical models and data, and research by Noll et al. (2018) and Justice et al. (2018) highlight the communicate power with which students and teachers alike think about statistical models.

As narratives describing events are constructed post-hoc, they cannot serve as a method for explicating students' in-the-moment thoughts. A retrospective self-constructed reflective narrative may represent what each individual has retained from the thinking episode. Thus, they may so provide a current state of thinking at the moment of reflection, but likely not all thinking that had occurred during the problem-solving task itself. The narrative may not include false paths and garden paths. It may only include what the person considers to be relevant, or true, or what they believe represents a coherent and consistent set of relationships among each action and interaction undertaken. It is well known that memories can be easily altered via suggestion, either by the self or by external factors (e.g., Loftus & Pickrell, 1995). Furthermore, reflections may represent students' retrospective reasons for the choices and moves that were made during the problem-solving episode, which may not be the same as their in-the-moment reasons.

However, retrospective narratives can still serve an effective role as a tool when considered as a cognitive instrument (Robinson & Hawpe, 1986). While not all of the thoughts students may have had during the problem-solving process will be evident in a self-constructed reflective narrative, their choice of what to include and what not to may be indicative of key relevant moments of their thinking as well as their statistical knowledge.

Therefore, this proposal takes as given, based on prior theory and evidence, that statistical narratives ubiquitously exist as a psychological construction during acts of statistical thinking and reasoning. Furthermore, it is assumed that they are produced through the integration of a problem context, prior statistical knowledge both of concepts and procedures, and the functionality of the specific statistical tools utilized during the thinking process. As a unified overarching scheme, statistical narratives make visible the logical structure of an individual's retrospective thinking, and while they may not fully reveal all aspects of the thinking process during the problem-solving task, may be predictive of an individual's thinking during tasks. Finally, incorporating traces of statistical thinking such as gaze paths and video records of task completion help to, at least partially, reconstruct in-the-moment narratives as may have occurred internally to each participant during task completion.

2.4.2 Defining Thinking

While researchers in the field of statistics education generally describe statistical thinking as 'thinking like a statistician', there is currently no consensus definition for statistical thinking (Le, 2017). Some researchers have attempted to define statistical thinking in terms of expected behavioral responses when interacting with statistical stimuli,

such as recognizing the omnipresence of variability or recognizing the need for data (e.g., Moore, 1990; Snee, 1993). Others have attempted to define statistical thinking in terms of expected cognitive processes, such as an interrogative cycle or investigative cycle (e.g., Wild & Pfannkuch, 1999). Exacerbating the problem is that the term statistical thinking is often used interchangeably with the term statistical reasoning, which is often defined in terms of more concrete actions such as explaining statistical procedures or interpreting statistical results (e.g., Ben-Zvi & Garfield, 2004).

More generally, researchers in the fields of cognition, metacognition, and epistemic cognition typically define thinking as a metacognitive process (Kitchner, 1983). For example, Moshman and Tarricone (2016) define thinking in relation to inferences, where inference is the generation of new knowledge, thinking is the metacognitive self-regulation of inferences, and reasoning is the epistemological self-regulation of thinking – "In thinking, one deliberately controls one's inferences on the basis of one's knowledge about inference in general and awareness of one's own inference" (p. 54).

Although metacognition and self-regulation are related as they are both types of cognitive control processes (Schunk, 2008), metacognition is typically defined in terms of conscious processes and the active coordination of strategies to accomplish a task (e.g., Howard et al., 2000). Conversely, self-regulation is typically defined in relation to specific actions that are undertaken, such as planning, revising, or checking (e.g., Baker & Brown, 1984). Furthermore, self-regulation is often defined as those regulatory responses to attention stimulated by one's environment, while metacognition is often defined as judgements or evaluations resulting from the mind of the individual as the initiator or trigger (Dinsmore et al., 2008).

Taken together, metacognitive self-regulation refers to individuals' decisions in choosing and monitoring appropriate actions, or put rather simply, making up one's mind about what to do (Bailin et al., 1999). There are generally three components to metacognitive self-regulation – planning, monitoring, and evaluating (Schraw & Moshman, 1995). Planning includes goal setting and the selection of appropriate strategies, monitoring includes self-checking on the progress of actions towards achieving one's goals, and evaluating includes revising one's goals and strategies when necessary (Schraw et al., 2006). In this way, metacognitive self-regulation is the link between goal setting and action, and the continual recalibration of actions to align with one's goals.

As thinking is inherently dependent on goal setting, one's purpose is thus central to thinking (Kaplan et al., 2009). Purpose is typically described as having three components: (1) perceived purpose for a specific task amidst a given scenario, (2) the identification of relevant self-processes, and (3) the actions perceived to be relevant to achieve the identified process for a particular purpose (Maehr, 1984). As perceived purpose is subjective to the individual, it is entirely possible if not probable that students interpret a task in different ways, perhaps even in different ways from those intended by task writers. They may thus adopt different purposes for engaging with the task, leading to different types of actions and thus metacognitive self-regulation (Kaplan et al., 2009; Winters et al., 2008). Therefore, planning as a component of thinking, and subsequently monitoring and evaluating the initial plan, is dependent on an individuals' initial perceptions of a task's purpose.

Acknowledging the variation in operational definitions of statistical thinking utilized by statistics education researchers as well as the prevailing approaches utilized by

researchers studying cognition and metacognition, this dissertation defines statistical thinking to be the metacognitive self-regulation of statistical inferences (and statistical inference as the generation of new statistical knowledge or the statistical generation of new knowledge). Put another way, statistical thinking is the management of taking statistical actions, or deciding what to do when faced with a statistical problem and monitoring these actions' progress.

2.4.3 Measuring Thinking

The lack of a single consensus definition for thinking has led to researchers utilizing a wide variety of tools to capture empirical evidence of thinking. A systematic review of research on thinking by Dinsmore et al. (2008) found that 73% of articles investigating self-regulation utilized self-report as their measure and 20% utilized observation. For those articles investigating metacognition, 24% utilized self-reports and 20% utilized observations. Additionally, 31% of papers reviewed utilized performance ratings, 12% utilized think-alouds, and 13% utilized interviews. One possible explanation of this variety is that metacognitive and self-regulatory practices are not always conscious or explicit choices and, therefore, may not be available to study participants to explicate and are thus difficult to capture (Winne & Perry, 2000).

As a result, many researchers are experimenting with creating new methods to investigate thinking. For example, Nicholas (2018) argues for the utilization of multiple videod events to provide a robust empirical basis for researchers' claims. In studying parents' thinking when reading to their children, Nicholas captured video recordings of the parent-child interaction, a post-hoc interview with the parent, and a delayed retrospective video-cued interview with the parent in which the parent was played selected clips from

the original video recording. The data produced by these different sources allowed Nicholas to analyze the phenomenon in greater detail than when using only one source alone.

Similarly, Dinsmore et al. (2008) argue that researchers have an opportunity to use subtle techniques to investigate thoughts, and to build tasks that not only will elicit metacognitive and self-regulatory awareness and reflection, but which will provide an opportunity for researchers to document subtle actions without intruding on the participants' thought processes.

One common tool used to study statistical thinking is the think aloud procedure. In a typical think aloud protocol, participants are encouraged to concurrently produce a verbal account of their thought processes during statistical acts. However, as participants produce verbalizations in the midst of problem solving, thinking aloud is an intrusive act that can change participants' thinking itself. For example, Baumann et al. (1992) found evidence that instructing students to think aloud led to an increase in their self-awareness and monitoring. While retrospective productions of a verbal account do not suffer from the problem of invasiveness, some researchers argue that the retrospective nature of these accounts do not accurately reflect the in-the-moment thought processes participants engaged in during problem solving. For example, participants with higher metacognitive skills may be able to more accurately retrospectively recall the strategies they used (Muijs & Bokhove, 2020). As a result, researchers have debated about the fidelity of think aloud interviews as a method to investigate thinking, with some arguing for it and some against (e.g., Ericsson & Simon, 1998; Smagorinsky, 1998).

An alternative to the individual think aloud protocol is group protocols, which allow researchers to capture the discourse and interaction between participants and make inferences about their thinking through methods from discursive psychology (Wiggins, 2016). While group tasks introduce social factors that must also be accounted for in analyses, they also ameliorate task discomfort, and make some types of decision making easier to observe (Schoenfeld, 1985). However, the social dynamics at play in group settings also cannot be ignored, as a dominant participant can skew the discussion. Especially if participants have not previously established a social dynamic, a spontaneously developing dynamic amidst a problem-solving task may further obscure the measurement and capture of evidence of each individual's thinking.

Due to the methodological concerns of individual and group think aloud strategies, a growing number of researchers argue for pure observational methods that can capture participants' thinking (e.g., Whitebread et al., 2009). These methods focus on capturing traces rather than think aloud statements. Traces are the observable signs of thinking that participants exhibit during a task (Dent & Koenka, 2015). These include actions such as underlining a passage, or even extended fixation and gaze towards a particular object. Some researchers have successfully utilized gaze-paths as recorded by eye tracking software to stimulate detailed verbal explanations of thinking in retrospective interviews (e.g., Cho et al., 2019; Hyrskykari et al., 2008).

While the utilization of traces avoids the invasive effect that think aloud procedures have, traces may not be able to fully capture thinking, and making inferences about thinking based only on these traces is an inherently uncertain venture for researchers. However, it is important to note that a similar critique can be made of think-aloud

protocols, in their inability to fully capture thinking, and that there is no consensus on which methods are better suited for which research questions. Some research has shown that thinking, cognition, and neural activity changes when participants are asked to think aloud (e.g., Durning et al., 2013; Fan et al., 2019), although the extent to which the think aloud protocol adulterates participants' naturalistic reasoning and thinking is debatable. For example, some researchers argue that thinking is inherently narrative, and therefore, thinking aloud is simply the external manifestation of a pre-existing internal monologue (see Cowan, 2019).

Perhaps more important than the method of data collection is the development of the task itself, as any observable action or behavior taken by a participant is inherently dependent upon the task they are interacting with (Maher & Sigler, 2014). Thus, any inferences made from data elicited by a task is dependent on the task's characteristics. Goldin (2000) argues that researchers must thoughtfully examine which characteristics are controllable, and which they have controlled for, as well as those they have not, to make valid and generalizable claims from interview data. Specifically, Goldin focuses on the content of the task, including the structure of the content presented, and its expected interaction with participants' cognitive structures.

Chapter 3: Method

To investigate the complex cognitive process of statistical testing and thinking about null models amidst the logic of statistical tests (after having completed initial instruction and training in statistics), to inform the design and modification of statistics curricula, and to support the burgeoning body of educational research on simulation-based pedagogies in statistics education, this dissertation focused on graduate students' thinking when conducting statistical tests. Furthermore, as the goal of graduate level statistics courses is to prepare graduate students for their own research, the focus of the conducted study was on graduate students' thinking seven months after the completion of a graduate level simulation-based introductory statistics course.

3.1 Research Questions and Study Purpose

The goal of this dissertation was to investigate the following questions:

(1) What is the nature of graduate students' thinking when conducting statistical tests?

(2) Do graduate students think about null models when conducting statistical tests, and if so, how?

To describe graduate students' thinking when conducting statistical tests and their thinking about null models in such tests, a case study approach was utilized to explore how each participant was thinking. Specifically, a multiple descriptive case (Merriam, 1988) based on multiple participants and multiple tasks was used to facilitate a description of how participants were thinking. Each case was a unique participant, and there were six cases.

The purpose of the study design was to elicit rich and detailed data regarding students' statistical thinking when conducting statistical tests. A small group of six graduate students participated in a set of structured interviews to generate detailed

information about their thought processes and conceptual understanding (see Figure 7). These results were intended to form an empirical record of these graduate students' statistical thinking in statistical tests.

Specifically, information from a semi-structured interview (the *Concept Mapping Task*) was intended to elicit empirical evidence of a conceptual model of students' thinking when conducting statistical tests (i.e., what is it that graduate students say that they will do when conducting a statistical test). Information from a task-based interview (the *Statistical Testing Task*) and retrospective interview (the *Video-Cued Interview*) was intended to provide empirical evidence of students' thinking when conducting statistical tests (i.e., what is it that graduate students do when conducting a statistical test). Information from a semi-structured interview (the *Statistical Testing Interview*) was intended to elicit evidence of students' thinking about null models and the extent to which this thinking depended on features of the software application utilized. All recruitment and study procedures were approved by the Institutional Review Board at the University of Minnesota (STUDY00016330).

3.2 Case Selection and Participants

Selecting cases required two considerations, how the participants would be selected and how the content for each task would be selected. The population of interest was graduate students who had completed an introductory statistics course utilizing a simulation-based approach to statistical inference. For expediency, the recruitment pool was limited to only students who had completed EPSY 5261, a master's level introductory statistics course at the University of Minnesota Twin Cities utilizing the Lock5 curriculum (see Appendix D for relevant excerpts of a EPSY 5261 course syllabus). While there will

be some differences in statistical thinking based on course format (in-person vs. online), the instructor's experience, and the participants' prior mathematics and statistics experiences and their grade in the course, the purpose of this study was simply to describe participants' statistical thinking at a level of detail hitherto unexplored for graduate students instructed with simulation-based approaches, and thus participants from all sections of EPSY 5261 were included in the eligible participant pool regardless of instructor or course format.

Participants were recruited from the pool of graduate students who had completed, within the past academic year as of the time of study recruitment, the introductory level simulation-based course in statistics offered at the University of Minnesota through the Department of Educational Psychology, EPSY 5261. This course generally followed the Lock5 curriculum and utilizes both simulation-based methods with either the StatKey or randomizeIt software application (delMas, 2021) as well as parametric-based methods (predominantly those based on the *t*-distribution) with the R software application (R Core Team, 2021). Two in-person sections plus an additional two online sections of this course were typically offered each fall semester with a total enrollment of approximately 50 total in-person students and 80 total online students. In spring semesters, one in-person section and one online section were typically offered with total enrollment of approximately 15 in-person students and 30 online students. One online section was typically offered each summer with a total enrollment of approximately 40 students. Email addresses of eligible participants were culled from course rosters, and eligible participants were contacted via email to participate in the study (see Appendix C).

A simple convenience sample of six students who had completed EPSY 5261 was used for this study. A total of 18 eligible participants volunteered, coincidentally all but one of whom were from in-person sections of EPSY 5261 in the Fall of 2021. Therefore, three participants were selected from each in-person section offered, with an attempt made by the research to balance the degree programs each student was a part of across each section. An initial screening verified that participants were familiar with both simulation-based methods utilizing StatKey or randomizeIt as well as parametric-based methods utilizing R for statistical testing. The selected participants were then provided with an electronic consent form and more information about the study (see Appendix E).

Participants were offered financial compensation for participation, in the form of an Amazon Gift Card, to incentivize study participation and retention. Participants were allowed to withdraw from the study at any time, in which case they would have been replaced in the study. However, no participants withdrew. Thus, each of the six participants completed the study in full and received $50.

Students of EPSY 5261 were exposed to two different software applications, StatKey or randomizeIt for simulation-based approaches to inference, and R for parametric equation-based approaches to inference. While idiosyncrasies in terms of the user interface as well as differences in the theoretical approaches they had been instructed to use with each software application varied, these differences were endemic to the nature of the EPSY 5261 course, and therefore, endemic to students' thinking about statistical tests. Including both software applications in this study provided an opportunity to triangulate each individual students' thinking about statistical tests. This facilitated capturing students' statistical thinking as a robust phenomenon, by situating statistical thinking in diverse

contexts and technologies. Therefore, participants completed one problem with simulation-based approaches through StatKey or randomizeIt and one problem with parametric-based approaches (i.e., the *t*-test) through R.

It is important to note that in the EPSY 5261 curriculum, StatKey or randomizeIt is utilized by students to conduct statistical tests in weeks 7–10 of the 15-week semester, while R is utilized by students to conduct statistical tests using parametric distributions in weeks 11–15 (see the Course Calendar included in Appendix D). Students are also taught how to use R to compute summary statistics to produce graphic displays of data in weeks 2–3 of the semester, and again practice these skills in weeks 11–15. Students utilize StatKey or randomizeIt to compute confidence intervals in weeks 4–6 of the course.

As the use of R for statistical tests came last, and the use of StatKey or randomizeIt came in the middle of the semester, there may have been a recency effect (Baddeley & Hitch, 1993) that could have affected the results of this study – students may have forgotten more about SBI and StatKey or randomizeIt than they had forgotten about parametric-based tests and R, simply as a function of the instructional sequence. However, this should not have affected their conceptual understanding if their concept of a null model was the same between both simulation-based methods in StatKey or randomizeIt and parametric-based methods in R.

3.3 Materials

The study consisted of a single session containing four separate components: a *concept mapping task*, a *statistical testing task*, a *statistical testing interview*, and a *video-cued interview*. All participants completed all four components. Additionally, at the end of the study, a detailed case history was obtained from each participant regarding their

experiences in EPSY 5261 as well as their experiences with statistics in general both before and after completing the course, including additional coursework, research, and other training.

3.3.1 Concept Mapping Task

The first task that participants completed was the *concept mapping task*. The purpose of the *concept mapping task* was to establish a baseline of students' conceptual model for the logic of statistical testing and the role null models play (i.e., what they thought they should do when conducting a statistical test). This task also served to establish participants' perceived purpose of engagement in conducting a statistical test, which would subsequently ground their planning, monitoring, and evaluating in statistical tests.

The *concept mapping task* took the form of a semi-structured guided interview, in which a series of open-ended questions were used by the researcher to elicit students' thoughts and internal logic (see Appendix F for the instructions explained by the researcher to the participants as well as the pre-prepared prompts that the researcher planned to use to elicit participants' thinking). Specifically, participants were asked to "draw a concept map for the logic of a statistical test" using pen and paper, with the researcher probing with follow-up questions to help participants add additional detail to their concept map.

Data collected from this task included the concept map drawn by participants, notes taken by the researcher during the task, an audio recording of the researcher and participant while completing the task, and a video recording of the concept map as it was being drawn by the participant.

3.3.2 Statistical Testing Task

After completing the *concept mapping task*, participants next completed the *statistical testing task*. The purpose of the *statistical testing task* was to observe students' statistical thinking as it was applied to conducting statistical tests using both simulation-based methods in StatKey or randomizeIt and parametric-based methods in R (i.e., what is it that they actually do when conducting a statistical test).

All participants completed two different problems as part of the *statistical testing task*, with participants utilizing simulation-based methods through StatKey or randomizeIt for one, and parametric-based methods through R for the other.

The two problems were based on contexts utilized in prior research by Biehler et al. (2015) and Brown (2021) to create the possibility of future cross-study comparisons. The first problem selected was the Verdienststrukturerhebung [Structure of Earnings Survey] (VSE) task utilized by Biehler et al. (2015). The VSE task is based on data collected by the German Statistisches Bundesamt [Statistics Bureau]. In the VSE task, students compare the monthly salaries of 861 women and men from 2006. Students first explore the data by comparing the distribution between two groups before they are prompted to conduct a null hypothesis test. The second problem selected was the Airplane Delays (AD) task utilized by Brown (2021). The AD task is based on data collected by the US Department of Transportation. In the AD task, students analyze the average delay time for Delta Airlines flights leaving the Minneapolis-St. Paul airport in 2019.

The two problems used in the *statistical testing task*, the VSE problem and the AD problem based on the VSE task of Biehler et al. (2015) and the AD task of Brown (2021) respectively, were constructed such that they were as similar as possible in terms of three criteria to control for potential confounding characteristics: both had a difference in means

as the parameter of interest, both utilized study designs with random sampling, and both were presented in a manner as homogenous as possible in terms of text characteristics (see Appendix G1 for the instructions the researcher provided to participants for this task, and Appendices G2 and G3 for the prompts provided to participants for each problem).

Data collected from the *statistical testing task* included notes taken by the researcher while the participant was completing the task, an audio recording of the researcher and participant while completing the task, a recording of the computer screen, and a recording of the participants' gaze while completing the task.

All participants completed the task utilizing an eye tracking apparatus that captured their gaze as they interacted with the software applications on their computer. The purpose of utilizing eye tracking was to capture a non-invasive trace of participants' thinking, particularly the way in which they monitored the task. Furthermore, it was thought that extended periods of gaze on a single object could provide an insight into the role that each piece of statistical information, and each software function, played in participants' statistical thinking. Additionally, this gaze recording would be part of the stimulus presented to participants in a retrospective interview. This evidence, when combined with the recording of participants' actions, their think aloud record, and the *video-cued interview*, were designed to provide a robust multi-modal perspective on participants' statistical thinking.

3.3.3 Statistical Testing Interview

The next task participants completed was the s*tatistical testing interview*. The purpose of the s*tatistical testing interview* was to probe students' thinking about null models across multiple statistical software applications and approaches to inference.

Therefore, the s*tatistical testing interview* focused on prompts that were likely to be highly discriminating in terms of students' thinking about null models.

The stimuli presented to participants in the s*tatistical testing interview* were related to the interpretation of results from statistical tests. Previous research has shown that students typically view the purpose of statistical testing in terms of the product of the test (i.e., a *p*-value; Justice et al., 2018; Noll et al., 2018b). A *p*-value is the result of a comparison between the null model and an observed sample statistic. Therefore, the task of explaining the logic of a statistical test or the story of a statistical test from statistical results was considered a prudent way to elicit students' conceptions of null models and the role they play in statistical testing.

The s*tatistical testing interview* thus presented participants with results from several different statistical tests, some of which were presented as output from parametric-based methods (i.e., *t*-tests) in R and some of which were presented as output from simulation-based methods (i.e., randomization tests) in StatKey (see Appendix H1 for the instructions the researcher provided to participants, and Appendix H2 for the stimuli presented to participants in this task). Two different statistical measures were utilized across the stimuli presented to participants – a difference between the means in two groups, and a single group mean. For every stimulus presented with simulation-based methods, there was a statistically isomorphic stimulus presented with parametric-based methods.

Data collected from this task included notes taken by the researcher while the participant was completing the task, an audio recording of the researcher and participant while completing the task, a recording of the computer screen, and a recording of the participants' gaze while completing the task.

3.3.4 Video-Cued Interview

After completing the s*tatistical testing interview,* participants then completed a v*ideo-cued interview*. The purpose of the v*ideo-cued interview* was to further probe students' statistical thinking. Specifically, the goal was to obtain a deeper understanding of their statistical thinking while conducting statistical testing tasks and about the role null models had in their thinking. The use of a video-cued and gaze-cued interview was intended to support a robust investigation of students' thought processes by supplementing the traces of participants' thinking obtained from the *concept mapping task,* the *statistical testing task,* and the *statistical testing interview*.

In the v*ideo-cued interview,* the researcher and the participant watched the gaze recording of the *statistical testing task* together. The participants were asked to comment on what they were thinking in each moment and the researcher and participant together could pause the video to comment on what they noticed from the recording to provide additional explanation when needed (see Appendix I).

Data collected from this task included notes taken by the researcher during the task and an audio recording of the researcher and participant while completing the task.

3.3.5 Pilot Testing

All initial materials, including tasks and interview prompts, were evaluated and revised based on a pilot test. Two participants were recruited for the pilot test in total. One participant was an expert statistician, with several years of training in statistics as well as in teaching statistics, including the EPSY 5261 course. The other participant had also received statistical training at an advanced level, well beyond the scope of EPSY 5261, but

had never been explicitly taught simulation-based inference in the manner taught in in EPSY 5261. These participants were compensated at equal rates as full study participants.

Feedback and results from this pilot study informed some revisions of the tasks and procedures. Specifically, it was decided that a concept map should be collected from participants both as the first task and as the last task, as participants may have recalled details while completing each task and may not actually remember enough about statistical testing to usefully create a concept map at the onset of the study.

Additionally, in the *statistical testing interview*, it was decided that the results from four of the ten stimuli would be intentionally edited to suppress *p*-values to force participants to think about the process of the statistical test, preventing them from thinking in a product-based manner. This manipulation was intended to specifically elicit participants' thinking about null models.

Finally, it was decided that in the *statistical testing task* all participants would complete the VSE problem using R, before completing the AD problem using StatKey or randomizeIt. This procedural change was intended to provide an opportunity for the researcher to assess participants' thinking about null models when using R without first priming them to think about null models (which may have occurred had they utilized simulation-based software applications first, in which the null model is graphically emphasized). If indeed participants' thinking about null models while using R were to change after participants utilized simulation-based approaches, it was determined that participants would have an opportunity to articulate this during the *statistical testing interview* and the *video-cued interview*. These changes were put in place before the six former EPSY 5261 students participated in the study.

3.4 Data Collection Procedure

All six participants met in-person with the researcher to complete all tasks. The participants were told that the study would take approximately 90 minutes to complete. Finally, participants were told not to prepare in any way or to review any materials or instruction from EPSY 5261 prior to participating in the study.

Upon meeting, the researcher first introduced the study, verbally reviewed the consent form and information sheet (see Appendix E), and obtained verbal consent from participants to record the meeting, including the participant's computer screen, any written notes, audio recording, video recording, and gaze recording utilizing eye-tracking software and equipment.

The researcher then introduced the *concept mapping task*, instructing participants to "draw a concept map for the logic of a statistical test", providing pen and paper to the participants to make a drawing (see Appendix F).

After completion of the *concept mapping task*, the eye-tracking apparatus was calibrated for each participant before participants were provided instructions for the completion of the *statistical testing task* (see Appendix G1). To make explicit their thinking, participants were prompted to explain their thoughts aloud and the researcher prompted participants to continue verbalizing their thoughts periodically throughout the task. All participants were instructed to use R to complete the VSE problem first. The original plan was to have participants utilize randomizeIt to complete the AD problem. However, for both the first and second study participant, who completed the study on the same day, there was a problem with running the randomizeIt application on the computer being utilized that was unable to be immediately debugged by the researcher. Therefore,

the researcher made the in-the-moment decision to utilize the StatKey application in order not to delay or extend the time needed for the participant to complete the study. While participants primarily utilized randomizeIt in EPSY 5261, the StatKey application, including images and explanations, were a part of the textbook utilized during the course (Lock et al., 2021). Additionally, the functionality of the randomizeIt application was largely based on the functionality of the StatKey application, but was designed to be internal to R. Therefore, the researcher determined that by providing some assistance to participants in terms of the procedural fluency required to utilize the StatKey application, utilizing StatKey to complete the AD problem would not interfere with the study of participants' thinking, especially given the functional similarity between StatKey and randomizeIt. Having made this determination, it was further determined to continue using StatKey for all study participants, to ensure similarity in procedures across all participants.

Upon completion of the *statistical testing task*, participants were given an opportunity to take a short break before the eye-tracking apparatus was re-calibrated. The researcher then introduced the *statistical testing interview* to further probe participants' thinking about null models and the role software applications played in participants' thinking. For each stimulus, participants were asked to tell the story of the test based on the results they were presented (see Appendix H1). The researcher also prompted participants to explain their thinking further when needed, especially to elicit descriptions about the nature of the null model that facilitates the comparison that produces a *p*-value. This was particularly true for the stimuli in which the *p*-value was intentionally suppressed from the output.

Upon completion of the s*tatistical testing interview*, the eye-tracking apparatus was removed and the participant and researcher together watched a recording of the participant's gaze and the computer screen from the *statistical testing task*. In this v*ideo-cued interview*, the researcher further probed the participants' thinking retrospectively in each moment through a semi-structured interview and informal probes (see Appendix I). These probes specifically focused on eliciting participants' reflections of their monitoring and evaluating, which were determined to be the more difficult aspects of thinking of which to acquire an in-the-moment empirical trace. Additionally, the researcher presented to participants a summary of the researcher's preliminary analysis and notes from the *statistical testing task* and asked the participants to either verify this preliminary analysis or to provide additional details to help provide the researcher a more complete understanding of the participant's thinking.

Finally, participants once more created a concept map for the logic of a statistical test before completing a case history. The case history interrogated participants in terms of their prior classroom training in statistics, their experiences with statistics professionally or as part of a research lab, and their perceptions and attitudes towards statistics. After this point, the participants were provided an opportunity to ask any questions about the study before the session officially ended.

3.5 Analysis Plan

This study utilized a multiple descriptive case study design, specifically one with an interpretivist stance (Merriam, 1988). The subsequent sections describe the assumptions made prior to analysis (i.e., the frameworks, propositions, stances, and definitions), as well as the procedures utilized to analyze the data collected from participants.

3.5.1 Assumptions

This study utilized a definition of thinking from the field of epistemic cognition, defining thinking as metacognitive self-regulation operationalized as planning, monitoring, and evaluating (Moshman & Tarricone, 2016). This is consistent with definitions utilized by statistics education researchers. For example, delMas (2004) defines a demonstration of statistical thinking as "a person who knows when and how to apply statistical knowledge and procedures" (p. 85).

It was assumed that graduate students with limited experience in statistics (i.e., only having completed a single introductory level statistics course) would struggle in thinking through statistical problems that deviate even slightly from the benchmark examples provided during instruction. This would be reflected in pauses and moments of silence, where they were metacognitively self-regulating (i.e., deciding what they should do next). It was also assumed that students' statistical reasoning (i.e., their epistemological self-regulation of thinking) would be weak, and they would not always be able to articulate why they should do something. However, it was further assumed that they would have a variety of tools that they would know how to use, and their negotiation of what they do and when they do it, in the absence of demonstrable statistical reasoning, could highlight how students interact with (1) data, (2) hypotheses, (3) the logic of statistical tests, and (4) the interaction between the statistical world and the real world.

To analyze data artifacts and make inferences about participants' thinking, this study utilized an interpretivist epistemological stance. Interpretivism takes as granted that knowledge is subjective and, in the context of case studies, seeks to explain phenomena as experienced by the participant, rather than some external 'objective' frame. Interpretivist

case studies are judged by the degree to which findings are consistent with participants' views (i.e., their credibility). The credibility of the findings of this study was designed to be supported by (1) providing participants an opportunity to comment on their own thinking in the *video-cued interview*, and (2) providing participants an opportunity to 'member-check' the researcher's analysis and provide additional comments.

As the object of this investigation was to describe each students' statistical thinking in statistical tests, and in particular their thinking about null models, the main foreground theory was the role of the null model in the logic of statistical tests. A historical analysis was used to identify the null model as the key component to the logic of statistical testing. The null model was defined as the expression of a (null) hypothesis in the form of a probability distribution, without which a statistical test cannot be performed. It is also important to note that simulation-based inference attempts to place the logic of statistical inference at its core and, therefore, attempts to place the null model at its core.

3.5.2 Analysis Procedures

To analyze data collected from each participant, this study utilized the constant comparative method with an inductive open coding process (Glaser & Strauss, 1967). Relevant codes were not a priori determined, but rather, stemmed from what was noticed in the data. However, inductively generated codes were ultimately organized into categories based on the definition of thinking utilized in this study. Specifically, codes were categorized based on whether they addressed planning, monitoring, or evaluating.

In the constant comparative method, one examines a single data artifact from a single case one at a time. Therefore, for each participant, analysis began by examining the data artifacts collected from the *concept mapping task*, the first task that participants

completed. Relevant moments were identified and documented in researcher analysis memos and assigned initial codes. Then, the analysis continued by moving on to data from the second task, the *statistical testing task*. Initially, only the *statistical testing task* was considered, with relevant moments documented in analysis memos and initial codes assigned for each relevant moment. Next, these memos and codes from the *statistical testing task* were compared to those identified from the *concept mapping task*. This process continued for the data from each successive task, and subsequently each participant. After the analysis of all cases was completed, the codes were compared across participants to ensure that relevant aspects and features of each participant's thinking were robustly described. Importantly, only the codes were compared across participants, and not the traces and descriptions of each participant's thinking. In this way, the analysis procedure was designed to roughly encapsulate line-by-line analyses as well as analyses meant to characterize participants' thinking in a manner corresponding to the first and second levels of qualitative data analysis as described by Simon (2019).

As the goal of this study was to describe participants' thinking and the logic they employ during statistical testing, four strategies were utilized to ensure that the data were interpreted in a credible manner. First, the design of the study employed participatory elements: the *video-cued interview* allowed participants to view the data they generated in the *statistical testing task* and provide an analysis of their own thinking. Second, the use of multiple types and sources of data served to triangulate participants' thinking, helping to ensure that empirically based interpretations of participants' thinking were not heavily influenced by quirks in the data collection or analysis processes nor quirks of a single modality or task. Third, participants were provided with an opportunity to member-check

analyses and descriptions of their thinking. Fourth, researcher memos were created throughout the data analysis process to document and divulge researcher biases that may have affected the analysis.

Data were used to support inferences that could credibly describe participants' thinking. These inferences were limited to the study participants – they were not meant to characterize all graduate students who had completed EPSY 5261, nor the efficacy of the EPSY 5261 curriculum. Furthermore, no causal inferences were made in terms of the reasons as to why participants may have thought in a certain way – the focus was on exploring and describing how participants thought.

To answer the first research question – What is the nature of graduate students' thinking when conducting statistical tests? – participants' thinking is described in terms of their planning, monitoring, and evaluating in statistical tests. The general definition of each as a distinct aspect of thinking was as follows – planning includes goal setting and the selection of appropriate strategies, monitoring includes self-checking on the progress of actions towards achieving one's goals, and evaluating includes revising one's goals and strategies when necessary (Schraw et al., 2006).

In the context of a statistical test, planning might take the form of the specification of what steps must be taken and in what order, monitoring might take the form of considering whether each step in the process is specified sufficiently and correctly, and evaluating might take the form of changing the plan for the test or questioning the monitoring process (see Table 3 for examples of codes from various sources and tasks that align with each of these components of thinking).

For example, one might make a plan to compute summary statistics for two groups, then monitor how the information generated supports the overall purpose at hand (i.e., the comparison of two groups to answer a research question), and subsequently evaluate whether additional computations are necessary or if the way in which the information is being connected to the overall purpose needs to be reconsidered. Each of these facets of thinking was described for each participant drawing on evidence from all sources of data.

To answer the second research question – Do graduate students think about null models when conducting statistical tests, and if so, how? – participants' thinking was described with a focus on how they thought about sampling variability, and especially how they thought about the simulation-based null model they encountered in SBI portions of the study. Once more, empirical traces supporting inferences about participants' thinking were drawn from all sources of data and all tasks.

Furthermore, to protect the identity of the participants, each participant was assigned a randomly generated pseudonym unrelated to their gender, racial/ethnic, or other sociocultural identity. These names were selected from a list of common names worldwide. This process was based on identifying monosyllabic or disyllabic names that were as gender, culturally, and linguistically neutral as possible. After a set of six such names were selected by the researcher, they were randomly assigned to participants using a random number generator.

Chapter 4: Results

This chapter presents the results relating to the research questions for each participant. As such, it is organized by participant. The section for each participant begins with a description of that participants' relevant background. After that, the results for each

of the research questions are presented. It is important to remember that participants were asked not to prepare prior to participating in this study.

4.1 Participant One – Jaci

Jaci took EPSY 5261 in the Fall of 2021, seven months before participating in the study. Jaci did not take any other statistics course between completing EPSY 5261 and participating in this study. However, Jaci did take a course on educational and psychological measurement in the Spring of 2022. To Jaci, the measurement course focused on concepts rather than statistical analyses, and Jaci saw the course as mostly unrelated to EPSY 5261. While an undergraduate student, Jaci also took a simulation-based introductory level course in statistics (based on the CATALST curriculum), although that was seven years prior to Jaci participating in this study, and Jaci said that they hardly remember anything from that class at all. Aside from the statistical software Jaci used in EPSY 5261, and the very limited use of R in the measurement course Jaci took, Jaci had not used any other statistical software.

Jaci very rarely did any statistical analysis for their own work or research, and only very rarely discussed statistics with colleagues. Those discussions typically took the form of evaluating whether various interventions were empirically sound, but the emphasis was typically placed on the study design, especially the sample size and the representativeness of the sample, rather than a thorough examination of the analysis and statistical methods. Important to remember is that Jaci was asked not to prepare prior to participating this study.

4.1.1 Jaci's Thinking When Conducting Statistical Tests

To Jaci, the purpose of a statistical test "is determining whether or not the null hypothesis or the alternative hypothesis is true" (Appendix P04-A, 00:09 – 00:16). Put

another way, the purpose of a statistical test is to determine "whether or not the difference observed in the sample would occur naturally" (Appendix P04-B2, 00:12 – 00:15). To Jaci, naturally occurring meant that "regardless of who you sampled you would likely get a very similar result" (Appendix P04-B1, 02:44 – 02:50).

Jaci primarily achieved this determination by generating a bootstrap dot plot from the data, and comparing a null hypothesis – typically that "no difference exists" (Appendix P04-A, 00:20 – 00:25) or that "no difference is observed" (Appendix P04-B2, 00:34 – 00:37) – to this distribution to determine whether or not the null hypothesis is likely to be true (see Figure 8). Jaci's determination of likeliness was based on an explicit specification of sampling variability, although not in the form of a null model. Instead, Jaci generated a distribution with the observed sample statistic at its center, i.e., a bootstrap dot plot. If that distribution was "centered very much around zero, this would be more proof that the null hypothesis is correct" (Appendix P04-A, 00:58 – 01:20). Whereas, if the distribution was not centered at zero, and if zero was in the tails of the distribution, "that would lead us to believe more that the alternative hypothesis would be the true answer to the research question" (Appendix P04-A, 01:32 – 02:18).

To Jaci, it was important to generate a large number of simulated trials when constructing the bootstrap dot plot. Jaci explained that doing so "provides more trials so that I can see what would likely happen under, um, whatever circumstances would occur based on the original sample" (Appendix P04-B2, 01:49 – 02:12). Furthermore, to Jaci, one purpose of generating more simulated trials was that "there is a certain number where it becomes more conclusive" (Appendix P04-D, 10:24 – 10:30). To Jaci, the logic of bootstrapping was that "each of these little points signifies a different hypothetical trial

based on the data in the original sample" (Appendix P04-D, 12:23 – 12:29). Furthermore, "all of the [simulated] data is going to be based around the original sample so the average is always going to be pretty similar, but it's basically just showing where the likelihood of each outcome winds up on the spectrum" (Appendix P04-D, 12:34 – 12:49).

4.1.1.1 Jaci's Planning of Statistical Tests

Jaci's plan for a statistical test, after having already acquired the data in hand, began with determining which test to utilize (Appendix P04-B1, 01:30 – 01:40). Specifically, Jaci noted that "There are different types of tests for finding 'is there a difference?' versus 'what is the difference?'" (Appendix P04-B1, 00:25 – 00:30).

Having determined which test to conduct, Jaci specified a null hypothesis. To Jaci, when "the research question is determining whether or not there is a difference … the null hypothesis would be that the true difference is equal to zero" (Appendix P04-D, 03:51 – 04:08).

Jaci then generated (or desired to generate, in the case of using R) a bootstrap dot plot, and through it obtain a 95% confidence interval estimate. When Jaci utilized R in the s*tatistical testing task,* after having generated output from the *t*-test, Jaci commented "should I be trying to make a graph of this?" (Appendix P04-B1, 08:28 – 08:33). When pressed by the researcher about what Jaci was hoping to see, Jaci explained that "I think seeing the difference itself plotted would help me get a sense for what the … for how stark the average difference actually is, under all possible trials" (Appendix P04-B1, 10:30 – 11:00). Thus it appears Jaci intended to make a bootstrap dot plot. To Jaci, it seems that seeing this bootstrap dot plot is a necessary step in determining whether a difference is significant. As Jaci further explained, "I can definitely see from running the *t*-test that there

is a difference, but it's whether or not the difference is significant … Because the computer science mean and the engineering mean, the computer scientists mean is lower and since both levels of the confidence interval are negative it means that under 95% of these circumstances, these trials that we could run, it would always fall in between those two values, well it would fall between those two values in 95% of the trials run. So I mean it seems likely that there is a difference" (Appendix P04-B1, 11:30 – 12:35). In other words, because the value of 0, which is the value of the parameter as specified by the null hypothesis, is not contained within the confidence interval, Jaci determined that there likely is a difference in the means of each group.

When utilizing StatKey, Jaci again desired to generate the bootstrap dot plot, but incorrectly interpreted the randomization dot plot based on the null hypothesis as the bootstrap dot plot based on the observed sample statistic. Thinking that a bootstrap dot plot and confidence interval were in hand, Jaci then generated a *p*-value. To determine whether or not the difference was significant, Jaci explained that "I mean I guess looking at the *p*-value would be the number one tell. And I think that a low *p*-value means that it's pretty, it is pretty likely that there is a significant difference" (Appendix P04-B1, 12:42 – 13:03).

4.1.1.2 Jaci's Monitoring of Statistical Tests

The primary manner in which Jaci monitored the statistical test (while using both R and StatKey) was through examining the upper and lower bounds of the confidence interval. For example, when completing the VSE problem in R, after looking at the results, specifically the confidence interval, Jaci commented, "okay so it seems like one of them [the group means] is lower on average since they [the upper and lower confidence bounds] are both negative" (Appendix P04-B1, 07:58 – 08:05). Examining both the upper and lower

bounds helped Jaci monitor the center of the distribution. In the *video-cued interview*, Jaci explained that while using StatKey in the AD problem, through looking at the tails of the randomization dot plot, then the middle, then again the tails, that "I think I was looking essentially to see like, if the, like what the different parameters were, to see whether or not they were truly fairly centered around zero, or if there was maybe a difference on either side" (Appendix P04-D, 08:02 – 08:25).

While Jaci's explanation on the surface may imply that Jaci was only interested in the location of the center of the bootstrap dot plot, Jaci actually did factor in the relative likelihood of various parameters based on the height of the dot plot. For example, in the *video-cued interview*, Jaci commented that, "looking at the distributions in the middle helped me get a sense for like, oh these are the most average outcomes" (Appendix P04-D, 09:28 – 9:35).

Additionally, while utilizing StatKey in the AD problem, Jaci focused on the mode of the randomization dot plot, noting that, "I am noticing that there are some points significantly higher, right underneath the null" (Appendix P04-B2, 02:46 – 02:54; see Figure 10). Looking at this mode, Jaci commented that "this is the most likely scenario that will ever occur" (3:58 – 04:07). However, this piece of information was not the only piece of information that Jaci used to monitor the steps of conducting the statistical test – Jaci also noted the center of the distribution, stating that "the mean is still pretty dang close to zero though" (Appendix P04-B2, 03:20 – 03:24).

While utilizing StatKey to complete the AD problem, Jaci also attempted to monitor the statistical test by comparing the distribution of the original sample to the distribution of the simulated trials (Appendix P04-D, 16:33 – 16:49). Specifically, Jaci was attempting

to pay attention to any differences in these two distributions, including differences in shape. For example, Jaci noted that the sample distribution was bimodal, while the distribution for one particular simulated trial was not (see Figure 9). However, Jaci did not appear to remember what to do with this information. Rather, Jaci appeared to simply be searching their memory for what distributions typically looked like and noticed anything that was different between the two distributions being compared.

4.1.1.3 Jaci's Evaluating of Statistical Tests

To Jaci, the bootstrap dot plot was core to the logic of a statistical test, and therefore, the main evaluation Jaci performed while conducting a statistical test was in ensuring that the bootstrap dot plot could be created. For example, after having generated results from a *t*-test in R as part of the VSE problem, and having drawn a conclusion based on the confidence interval and *p*-value, Jaci still asked, "Should I be trying to make a graph of this?" (Appendix P04-B1, 08:28 – 08:33). When prompted to explain what a graph might add in terms of answering the research question, Jaci explained that "I think seeing the difference itself plotted would help me get a sense for what the … for how stark the average difference actually is, under all possible trials" (Appendix P04-B1, 10:30 – 11:00).

4.1.2 Jaci's Thinking about Null Models

Although Jaci did think about a sampling distribution for a sample statistic as a key component of statistical testing, it was not a null model that Jaci thought about. Jaci did not describe sampling distributions under a null hypothesis. Instead, Jaci described the sampling distribution as a distribution where "all of the [simulated] data is going to be based around the original sample so the average is always going to be pretty similar, but it's basically just showing where the likelihood of each outcome winds up on the spectrum"

(Appendix P04-D, 12:34 – 12:49). Jaci was describing a bootstrap dot plot, which might be construed as something akin to a likelihood function, but not as a null model.

Jaci did, however, seem to think about sampling variability as an important part of statistical testing. For Jaci, sampling variability was whether or not "regardless of who you sampled you would likely get a very similar result" (Appendix P04-B1, 02:44 – 02:50). In this way, sampling variability was the key determinant of significance to Jaci. Expressing sampling variability would, in Jaci's own words, "help me get a sense for what the … for how stark the average difference actually is, under all possible trials" (Appendix P04-B1, 10:30 – 11:00).

To Jaci, this manifestation of outcomes under all possible trials "provides more trials so that I can see what would likely happen under, um, whatever circumstances would occur based on the original sample" (Appendix P04-B2, 01:49 – 02:12). These circumstances helped show Jaci "where the likelihood of each outcome winds up on the spectrum" (Appendix P04-D, 12:34 – 12:49). This variability of circumstances and likelihood of each outcome appeared to serve as something of a stress test for Jaci, in terms of the robustness of the observed sample statistic. Indeed, Jaci commented on the exhaustiveness of generating simulated trials, stating that "I remember there is a certain number where it becomes more conclusive" (Appendix P04-D, 10:24 – 10:30). To Jaci, it appeared that if, based on the observed data, over the course of many simulations there were no (or very few) simulated outcomes of a particular parameter, then that parameter was unlikely to be true. For example, as Jaci described in the *concept mapping task*, "if [the bootstrap dot plot] was something more like, [draws a bell curve] … let's say zero was right here [off center] … and so that would lead us to believe more that the alternate

hypothesis would be the true answer to the research question" (Appendix P04-A, 01:32 – 02:18). Jaci did however still think in terms of the middle 95% of the bootstrap dot plot, equivalent to the confidence interval approach to hypothesis testing, and functionally no different from a '$p < .05$' decision rule. For example, as Jaci described in the *statistical testing task*, "under 95% of these circumstances, these trials that we could run, it [the null parameter] would always fall in between those two values, well it would fall between those two values in 95% of the trials run. So I mean it seems likely that there is a difference [because it doesn't fall between these values]" (Appendix P04-B1, 11:30 – 12:35). However, Jaci did not strictly follow a decision rule when interpreting results. As described above, Jaci also made consideration for the relative likelihood of parameters.

4.2 Participant Two – Kei

Kei took EPSY 5261 in the Fall of 2021, seven months before participating in the study. Kei did not take any other statistics course between completing EPSY 5261 and participating in this study, nor did Kei take any statistics or quantitative methods courses since completing their undergraduate degree. Kei had used SPSS while an undergraduate student, but had not used SPSS since, nor any other statistical software outside of EPSY 5261.

Kei had not had to do their own statistical analyses for any of their projects. Even in Kei's research lab, all quantitative work was typically contracted out to a statistical expert, and the lab rarely discussed statistical methods and analysis. Kei had read journal articles that employ quantitative methods, and did focus on reading the methods section, but admitted to not always knowing what they were reading. As with all participants, Kei

was told not to prepare prior to participating in this study, and acknowledged that they truly did not prepare, despite wanting to.

4.2.1 Kei's Thinking When Conducting Statistical Tests

Kei's thinking seemed to be highly based on the manner in which Kei was taught in EPSY 5261. This was most evident in Kei's planning, which seemed to be the result of a well-rehearsed rote procedural fluency. However, Kei described being overloaded with information when interacting with statistical software while completing this study. Kei also admitted that the results felt like a foreign language, and that they knew that certain output needed to be generated but did not know what to do with the output once it was generated. Kei commented that "the thing is, I wasn't, like, thinking … I was just trying to plug [stuff] in" (Appendix P03-D, 05:26 – 05:34) and that "when I do stats, it's very, like, procedural" (Appendix P03-B2, 03:15 – 03:33). Nevertheless, Kei's high procedural fluency seemed to help Kei focus on the process of the test, rather than only the product of the test (i.e., *p*-values).

4.2.1.1 Kei's Planning of Statistical Tests

While Kei's planning did prima facie differ between each of the statistical software tools Kei used, there did appear to be a general approach to Kei's planning that was consistent across all software. As seen in Table 4, Kei first began by importing a dataset into the software and checking the data to ensure the import occurred successfully. Next, Kei examined each group separately. Kei was particularly explicit about the fact that each group should be examined separately, and commented that, "I need to see each individual graph … and that's how I was taught, I don't know any other way" (Appendix P03-D, 12:30 – 12:36).

After examining each group individually, Kei then compared the two groups by computing the difference in means. This was quickly followed by Kei interpreting the *p*-value and 95% confidence interval estimate. Interesting, the choice of interpreting *p*-values or confidence intervals was dependent on which statistical software Kei was using – Kei focused on confidence interval estimates while using R but focused on *p*-values while using StatKey. Kei explained that they focused on the *p*-value in StatKey for its ease, commenting that "I feel like there's so much stuff to look at in that [StatKey] graph that I just shut down … but here [in R] it's very clear" (Appendix P03-C, 06:25 – 06:40). This acceptance of both confidence intervals and *p*-values as tools to conduct a statistical test is also reflected in Kei's concept map for the logic of a statistical test (see Figure 11) in which Kei gives equal prominence to both. However, Kei's explanation also suggests that Kei would have preferred to think about confidence intervals in StatKey, but found it too confusing to do so, and thus deferred to the *p*-value.

4.2.1.2 Kei's Monitoring of Statistical Tests

Kei's monitoring seemed to be entirely based on Kei's visual memory and recognition of familiar output. For example, when commenting upon how Kei might monitor whether 'checking the data' had gone well, Kei stated, "I check the data, to see if it's … I don't really know, I would know if I saw it" (Appendix P03-B1, 03:01 – 03:12). Kei also commented that "the thing is, I wasn't, like, thinking … I was just trying to plug [stuff] in" (Appendix P03-D, 05:26 – 05:34). These comments suggest that Kei's monitoring was based on a well-rehearsed routine, perhaps one practiced and reinforced by the activities and assessments in EPSY 5261. Furthermore, it suggests that Kei had some recollection of what things should look like based on this experience and training in EPSY

5261, but that this monitoring was rote in nature. Kei tacitly acknowledged this. In the *video-cued interview*, the researcher commented that after generating summary statistics in R, the gaze recording indicated that Kei barely looked at the output generated, to which Kei responded "at this point, that's like a foreign language … I was like I don't even know what to do with that, like, I don't even know what that means" (Appendix P03-D, 09:48 – 10:02).

4.2.1.3 Kei's Evaluating of Statistical Tests

Many times, Kei commented upon not really knowing what to do with information being generated, even though Kei knew that specific pieces of information needed to be generated. However, while Kei was not always able to exactly specify in which way their plan should be amended, Kei eventually was able to describe what they had wanted to do, especially with the help of the researcher.

For example, in the *video-cued interview*, Kei commented that "I remember doing this code [in R] enough, where like you're always doing, thematically you're doing, or you're inserting whatever, like, the theme of the thing" (Appendix P03-D, 07:12 – 07:30). In this explanation, Kei gave a clue that the theme of the task, in this case a comparison of salaries between majors, was the guiding light as Kei thought about conducting the statistical test.

The importance of the theme of the task was also reflected in how Kei thought about the graphs they were generating. For example, as reflected in the gaze recordings, and as Kei commented upon in the *video-cued interview*, "… originally I was looking in the middle, but then I was like, 'which salary makes more?' … because then I was thinking I need to see the ends, or like see if there's a lot of people towards the end" (Appendix P03-

D, 18:20 – 1843). The theme of the task – the comparison of the salaries between groups – is seemingly what guided Kei to shift their planning and change the way they were thinking about the statistical output.

Similarly, the importance of the theme was again reflected in the *statistical testing task,* where Kei, after trying to interpret a histogram graph they generated as part of the VSE problem, looked back at the research question and then stated, "wait a minute, I don't know, I feel like I need to do the other thing … umm, I want to find the average … I think I want to find the specific, I don't know, I don't really … cuz like 'is there a difference in the average salary?', so I'd want to find the average" (Appendix P03-B1, 16:00 – 16:38). Thus, Kei's evaluation of their initial plan for which steps should be conducted was seemingly based not on the output Kei was generating or the result of any one of the specific steps Kei was taking, but rather by the theme of the task.

4.2.2 Kei's Thinking about Null Models

It appears that Kei did think of the null model when conducting and interpreting results from statistical tests based on SBI, albeit in a hesitant manner. Specifically, Kei related the null hypothesis to the center of the distribution and recalled that the tails of the null model were an important part of the testing process.

As seen in the *statistical testing interview*, Kei related the center of the randomization dot plot being zero to the null hypothesis. After looking at the center of the randomization dot plot for the question 'Is average commute time in Atlanta and St Louis the same?' (see Figure 12), Kei commented, "Oh! I guess the null is zero." (Appendix P03-C, 01:08 – 01:12).

A few moments later, when answering the question 'Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?', Kei noticed that the center of the randomization dot plot was 28.6 – the value specified by the null hypothesis for the test for a single mean – and paused to reflect on what this meant (Appendix P03-C, 02:20 – 05:28). Kei spent several moments thinking about the problem, but eventually did seem to relate the null hypothesis value of 28.6 to the center of the randomization dot plot (see Figure 13). Kei subsequently drew the conclusion that there was a difference, rejecting the null hypothesis, because the *p*-value was in the tail of the distribution (see Figure 14). During the *statistical testing task*, Kei also commented on the tails of the distribution, albeit not fully remembering exactly why they were important, stating "I feel like I do something with the tails" (Appendix P02-B2, 06:33 – 06:35).

However, it is unclear whether Kei recalled that the null model is based on the underlying data generating process. In the *statistical testing task*, Kei commented that part of the test required generating samples but did not remember why this had to be done (Appendix P03-B2, 04:50 - 05:35). Kei's concept map for the logic of a statistical test (see Figure 11) did reflect a data generating process, but there was no empirical trace of Kei thinking about a data generating process while conducting or interpreting results from the statistical tests in either the *statistical testing task* or the *statistical testing interview*. Therefore, there is no evidence to support the claim that Kei did think about the null model as based on a data generating process in conjunction with a null hypothesis.

However, it must be noted that Kei did not mention the null hypothesis at all in the *statistical testing interview* when examining R output, instead commenting that "I feel like I can just think about it" (Appendix P03-C, 06:16 – 06:19), answering the research question

based on the confidence interval. Null hypotheses were also conspicuously missing from Kei's plan for conducting a statistical test (see Table 4). Thus, while it appears that Kei did think about the null model in relation to the null hypothesis, Kei may have only somewhat understood and been thinking about the null model's role in statistical testing.

4.3 Participant Three – Chau

Chau took EPSY 5261 in the Fall 2021 semester. Chau then subsequently took EPSY 5262 – Intermediate Statistical Methods, in the Spring 2022 semester. EPSY 5262 is a follow-on course to EPSY 5261 offered through the Department of Educational Psychology at the University of Minnesota Twin Cities which focuses on statistical models, particularly within the multiple linear regression framework. Furthermore, Chau had previously taken two statistics courses prior to completing EPSY 5261, one at the undergraduate level and one at the master's level, but both at the introductory level, and both at different institutions. Chau explained that they did not really remember much from those previous courses aside from *t*-tests, and also felt that they did not really understand statistics until taking EPSY 5261.

In terms of software, Chau considered themselves an "R newbie". Chau typically had used SPSS to conduct analysis, both as part of classes as well as outside of classes. However, Chau explained that whatever software was being used, the important part to them was the interpretation of results, something that Chau felt comfortable doing regardless of the statistical software used.

While Chau's research was mainly qualitative in nature, Chau had experience conducting statistical analyses, having served as a co-author on over twenty papers. Furthermore, Chau was the main analyst for some of these papers. These papers typically

included the computation of basic descriptive statistics. In terms of statistical inference, these papers primarily focused on the use of *t*-tests, especially the two-sample *t*-test and the paired *t*-test. However, it is important to note that Chau completed the majority of these papers' analyses before taking EPSY 5261. As with all other participants, Chau was instructed not to prepare or review any materials prior to participating in this study.

4.3.1 Chau's Thinking When Conducting Statistical Tests

To Chau, the purpose of a statistical test was to determine whether there existed a difference between groups (or between a single group and a hypothesized parameter). This was achieved by examining a *p*-value, and if that *p*-value was less than .05, then Chau determined there was a difference. In general, Chau's thinking, specifically in terms of Chau's planning, was remarkably consistent across tasks. However, when pushed by the researcher to explain their reasoning (even though measuring students' statistical reasoning was outside of the scope of the research questions of this study), Chau seemed to struggle to explain why some steps needed to be done as part of a statistical test, only knowing firmly that they must be done. Quite simply, Chau stated that if one does not do a statistical test then "we cannot say its [the difference is] statistically different" (see Appendix P06-A, 08:00).

4.3.1.1 Chau's Planning of Statistical Tests

Chau's planning is best evidenced by the concept map for the logic of a statistical test that Chau created (see Figure 15), which Chau also followed quite faithfully during the *Statistical Testing Task*. As seen in Chau's concept map, a statistical test started with Chau thinking about a research question, which was instrumental to Chau's planning of the statistical test, as it provided the grounding context (see Appendix P06-A, 04:00). Chau

saw the research question as specifying the type of test that should be conducted, whether it was a one-sample *t*-test, a two-sample *t*-test, or some other test. It also specified the significance level that Chau would use in interpreting the *p*-value, which Chau noted was conventionally .05 in the social sciences. Finally, Chau saw the research question as specifying whether the study was observational or experimental, which had implications for the analyses that Chau would conduct (see Appendix P06-B1, 01:00).

After the research question, Chau then thought about the sampling strategy. Ensuring that the sample was randomly selected was very important to Chau, although the reasons why are somewhat unclear from the evidence collected. Chau knew that a sample being randomly selected was one of the assumptions that must be met for a *t*-test, but did not articulate why this was an important assumption.

Next, Chau explored the data, by first opening the excel file containing the data (see Appendix P06-D, 01:55 – 02:10), and then by computing the sample mean and the sample standard deviation (Appendix P06-B1, 07:50 – 08:00). Chau generated both summary statistics as well as histograms to examine the data.

Chau then began to move towards conducting the *t*-test by verifying several of the test assumptions, including the normality of the data, and in general by thinking about the sample distribution, the sample size, and the sampling strategy. Chau then conducted the *t*-test, extracting a *p*-value. To Chau, the *p*-value was the most important statistic. Chau stated that "to answer the question, we use only the *p*-value" (Appendix P06-A, 06:54 – 06:58). To Chau, the *p*-value was what would show whether there was a significant difference or not. After Chau had determined that there was a difference, Chau then

generated a confidence interval to obtain the estimated difference between groups (Appendix P06-A, 07:12 – 07:17; Appendix P06-B1, 12:40).

4.3.1.2 Chau's Monitoring of Statistical Tests

As with Chau's planning, Chau's monitoring was also remarkably lucid, especially when Chau was utilizing R. Generally speaking, Chau made several comments indicating what information Chau was monitoring. For example, as Chau read the prompt for the VSE problem, Chau commented on the study design, noting that "it's an observational study" (Appendix P06-B1, 01:00). Similarly, as Chau created summary statistics later on while working on the VSE problem, Chau monitored which statistics were being generated, commenting, "so I have all the data, median, mean, and then quartile, for both of the groups, standard deviation" (Appendix P06-B1, 07:50 – 08:00; see also Appendix P06-D, 04:50 – 05:00). After creating histograms for the VSE problem, Chau, in the *video-cued interview*, commented that they were mainly looking at the shape of the distribution, with special attention towards noticing any skew (Appendix P06-D, 06:25 – 06:35). Chau also commented on remembering that when the sample size was greater than 30, they did not even need to check the shape of the distribution and that it was the summary statistics that were the important pieces of information to think about (Appendix P06-D, 06:35 – 07:35). Finally, having generated results from a *t*-test in R, Chau acknowledged that they were mainly looking at the test statistic and *p*-value, and that these were the most important parts of the output of the *t*-test (Appendix P06-D, 09:55 – 10:05).

When using StatKey, Chau also commented on what pieces of information were important to monitor, albeit not live during the *statistical testing task* but rather retrospectively in the *video-cued interview*. For example, Chau commented that after

generating simulated trials, Chau looked at the center of the randomization dot plot and then also the mean of all of the simulated trials (Appendix P06-D, 08:00 – 08:15). Chau also commented that they usually looked for the confidence interval and the *p*-value as well (Appendix P06-D, 08:45 – 08:55).

Generally speaking, in both R and StatKey, Chau's monitoring seemed to largely hinge upon the pieces of information needed to verify the assumptions of the *t*-test, as the goal of Chau's plan was to be able to conduct such a test to obtain a *p*-value and confidence interval.

4.3.1.3 Chau's Evaluating of Statistical Tests

Chau's evaluating seemed to stem from checking the assumptions for conducting a statistical test, as Chau specified in the *concept mapping task*. Particularly, Chau focused on the study design, specifically whether the study was observational or experimental, and additionally focused on the *t*-test assumptions such as the normality of the sample distributions or the sample size of each group.

However, one curious episode occurred as Chau was completing the VSE problem. Chau had generated summary statistics, a histogram, and verified *t*-test assumptions such as the normality of the data. However, as Chau was typing the function in R to conduct the *t*-test, Chau paused, and then stated, "sorry, I cannot do that because it's not an experimental study, right? So I cannot do the *t*-test here" (Appendix P06-B1, 08:50 – 09:00). Chau then proceeded to explain why, stating that "because I need an independent and dependent variable, but there is no independent and dependent variable here, so I just compare between two groups. So when we compare two groups, when we use the *t*-test, we have the treatment and the experimental group, and then also the treatment group and

then also the control group, but there is no control here. So we just use, they have the same coefficient here, so we just need to use the standard deviation actually to compare the difference, in this case" (Appendix P06-B1, 09:20 – 10:04). Chau then revised their plan to utilize simulation-based software tools to conduct bootstrap resampling to generate a confidence interval (see Appendix P06-B1, 12:40).

This is interesting for two reasons. First, Chau utilized the 95% confidence interval estimate to determine whether or not there was a statistically significant difference, based on whether 0 was contained within the interval. This implies that Chau did not think that the bootstrap dot plot and the *t*-distribution were equivalent methods for expressing the sampling distribution for this question. Secondly, when completing the AD problem, Chau commented that the context was the same as in the VSE problem, yet for the AD problem, Chau stated that they wanted to conduct the *t*-test in R. Why, if the VSE and AD problems were structurally similar, did Chau determine that the *t*-test could not be done for the VSE problem but could be done for the AD problem? The answer to this question is unclear based on the data collected.

4.3.2 Chau's Thinking about Null Models

It did not appear that null models played a large role, or perhaps any role, in Chau's thinking about statistical tests. In the *statistical testing interview*, Chau almost exclusively focused on whether the *p*-value was less than .05, even when provided output from simulation-based software showing the null model. Similarly, while completing the AD problem using StatKey, Chau did not seem to connect the center of the randomization dot plot to the null hypothesis, instead drawing the conclusion that there was no difference between the groups as the distribution was centered on zero. Thus, it appears that Chau

thought the randomization dot plot was actually a bootstrap dot plot. Furthermore, it appears that Chau did not explicitly think about null models when thinking about statistical tests, nor about the specific manner in which null hypotheses produced probability distributions through which the *p*-values were generated.

4.3.2.1 Did Chau Think about Sampling Variability?

It appeared that Chau did understand that considerations for sampling variability must be made, and were done so through conducting a statistical test, although it seemed that Chau did not understand exactly how these considerations played a role and were achieved in statistical testing. Thus it appears that Chau did not explicitly think about sampling variability while thinking about statistics tests. During the *concept mapping task*, when asked why one bothers with conducting a statistical test, Chau answered that without doing a test, "we cannot say it [a difference between two groups] is statistically different" (Appendix P06-A, 7:47 – 7:50). When asked to explain what 'statistically different' means, Chau explained that "statistically different means it's kind of, you know, we look for, that it likely happens for the sample, likely random sample and unlikely random sample happens to our population. To test that our randomized sample can represent our population, without this kind of statistical analysis, so, we can see only [the] difference between means. We can see the difference, but how much the difference [is] we cannot know, unless we do this kind of statistical analysis. We cannot know how much [it is] statistical[ly] different, only the means. Because the mean is only [the] calculation of the average, but how much [it is] different we cannot know. But then [with the test] we can prove, we can reject a null hypothesis, or we support [the] alternate hypothesis" (Appendix P06-A, 7:52 – 9:12).

On the surface, Chau's answer seems somewhat tautological. However, Chau's comment about "likely random sample and unlikely random sample happens to our population" might be an indication that Chau was thinking about sampling variability in terms of the sampling distribution under the null hypothesis as specifying what possible sample statistics were probable under the null hypothesis and which were not. However, Chau did not seem to demonstrate this thinking during the *statistical testing task*, failing to recall that the randomization dot plot is centered on the null hypothesis parameter, and that the dot plot is the exact specification of which sample statistics were likely and which were unlikely (i.e., the randomization dot plot was the null model). However, the simplest and perhaps best explanation is that Chau did not quite remember exactly what role sampling variability played in statistical tests, perhaps only remembering that it did play a role in some way.

4.4 Participant Four – Tal

Tal took EPSY 5261 in the Fall of 2021, seven months before participating in the study. Tal did not take any other statistics course between completing EPSY 5261 and participating in this study. However, Tal had a previous master's degree, which Tal had earned over a decade before completing EPSY 5261. While completing that master's degree, Tal took two basic statistics courses, the first of which was a traditional introductory course focusing on descriptive summary statistics as well as inferential statistics such as confidence intervals and hypothesis tests, and the second of which focused on general linear regression models.

In previous jobs held by Tal, Tal had often conducted quantitative analyses, but almost entirely in the form of computing summary statistics. Tal regularly would compute

frequencies, percentages, and averages, but almost never conducted *t*-tests, or any other statistical tests. In Tal's more recent research work, Tal similarly focused on frequencies, especially in analyzing survey responses. Tal rarely used regression or saw anyone else in their field use it, as regression was typically considered an advanced method. Similarly rare was Tal's or Tal's colleagues' use of formal statistical tests.

In terms of software, Tal primarily had used SPSS, but only had done so with their advisor and on a single research project. Other than SPSS, Tal had not used statistical software, nor had Tal ever seen R outside of EPSY 5261. Important to remember is that Tal was asked not to prepare for this study.

4.4.1 Tal's Thinking When Conducting Statistical Tests

For Tal, the fundamental purpose of a statistical test was to confirm a theory. Tal thought about null hypotheses (and research questions) as coming from theories, and when a null hypothesis was rejected, Tal thought that this indicated a potential problem with a theory. As a result, Tal would subsequently embark on a search for multiple possible other theories that might be sensible.

To Tal, this process of testing was iterative. Tal thought of the process as iterating through multiple rounds of evaluations of research questions and results until arriving at an understanding of the phenomena being studied – "If it [the result] doesn't make sense I have to go back to theory, even tweak the research question if necessary. Pretty much I have to do it again, I have to analyze the result again. And I keep doing that forever, until the results are not one that doesn't make sense. I do that until I get sensible results" (Appendix P02-A, 03:44 – 04:07).

Null hypotheses to Tal were ideally confirmed – "so if the null hypothesis is true, yes, so whatever I was thinking back when I started in my research question, my gut instinct was right. So whatever I was thinking, the data proved that whatever I was thinking, whatever theory I was thinking exists, it is true" (Appendix P02-A, 04:47 – 05:09). However, based on Tal's thinking in the *statistical testing task* and especially the *statistical testing interview*, it is unclear to what extent Tal distinguished between research questions and null hypotheses – it may be that to Tal, research questions were ideally confirmed while 'no effect' hypotheses were ideally rejected.

Alternate hypotheses to Tal were "just the opposite of the null hypotheses" (Appendix P02-A, 06:20 – 06:22), and there were always implied alternate hypotheses. However, despite the existence of alternate hypotheses, Tal's thinking was not a decision-theoretic, nor did Tal think about the purpose of the test as a choice between competing hypotheses. Indeed, Tal correctly explained that "if the null hypothesis is not true, that doesn't mean that the alternate hypothesis is true. It [the alternate hypothesis] can still be false" (Appendix P02-A, 06:59 – 07:07). To Tal, the test, through the analysis of evidence, was about confirming or disconfirming a single candidate theory, amidst the larger process of considering multiple theories, individually striking non-sensible ones until a sensible arriving at a theory that can be deemed sensible based on the observed evidence from the world (see Appendix P02-A, 08:00).

4.4.1.1 Tal's Planning of Statistical Tests

Tal's plan for conducting a statistical test, as Tal described when creating a concept map (see Figure 16), started with a theory. This theory then in turn informed a research question, which contextualized and applied the theory to a specific phenomenon. It appears

that to Tal, research questions were not really interrogative, but rather, laden with specific theory-based claims about a specific context.

Having determined a research question, Tal then determined which statistical method would be used to try to answer the research question, including how the data would be collected, and which statistics would be used to measure which characteristics. To Tal, this was one of the most important steps. As such, once the data was collected, it also had to be evaluated for quality. This also guarded against poor data quality as a cause for non-sensible results.

At this point, Tal then entered the "software world". As Tal noted in the *statistical testing task*, "I want to look at the dataset [first]. I usually do that" (Appendix P02-B1, 01:26 – 01:31). One important part of Tal's plan centered on the computation of confidence intervals. According to Tal, the 95% confidence interval "is how confident that you are that that information [the null hypothesis] is within the 95% interval of whatever data you have" (Appendix P02-B1, 08:15 – 08:20). Based on this explanation, Tal thought about sampling distributions as based on the data (as is the case with a bootstrap dot plot), and not as based on the null hypothesis (as is the case with a null model).

The key step in Tal's plan was a consideration of the "sensibility" of the data and the results – "I have to make sure it makes sense" (Appendix P02-A, 02:18 – 02:20). The results had to be contextualized, and Tal ultimately determined whether the results made sense or not against the research question and the grounding theory.

4.4.1.2 Tal's Monitoring of Statistical Tests

Tal's monitoring was most clearly evident in the way Tal thought about the nature of the observed data that would subsequently be a part of the statistical test. While

completing the VSE problem, having generated summary statistics for the salary for each group, Tal systematically scanned the output for the computer science group first, before scanning the output for the engineering group and making comparisons between the groups. Tal made these comparisons one-by-one for each statistic reported by R, stating "for engineers, the minimum is way higher, the median is a little higher, maximum [is] pretty much the same, mean for engineers [is] a little bit the same, standard deviation for engineers [is] a little bit lower. Uh, without really any further tests, my gut feeling, is, engineers have a lower range, computer science have a higher range. Uh, both have like pretty much the same top salary, and the medians, both medians are a little bit the same. If we're thinking in terms of percentage, … it's like a 10% difference in median, not 10, less than 10, like 7% difference in terms of the median. The mean is like 4% different. So without doing an actual statistical test, I would say that engineers on average earn a little bit more than computer scientists, and they have a lower range" (Appendix P02-B1, 04:18 – 05:56).

It is interesting to note that Tal mentally computed the percentage difference in the median and mean salaries between groups, in effect computing an effect size, to facilitate Tal's monitoring of the comparison between groups. It seemed that this statistic, especially the 4% difference in means, was an important piece of evidence in Tal's thinking. It is also interesting that Tal commented upon the range, noting that the range in engineers' salary was less than the range in computer scientists. This is an indication that Tal was comfortable thinking about a distribution. Similarly, Tal noted in the *video-cued interview* that looking at the five number summary helped Tal visualize the distribution for each

group, and that they were trying to picture the distribution (Appendix P02-C, 05:35 – 06:10).

In the *statistical testing task*, Tal seemed to monitor the determination of their initial plan by tracking both the comparison of the estimated difference in means against the null hypothesis as well as the confidence interval estimate and bootstrap dot plot, especially the tails of the distribution. Although Tal incorrectly thought they were looking at a bootstrap dot plot when they were actually looking at a randomization dot plot, Tal explained that, "What it is giving me is each individual 5000 sample data, of the average difference delay time from Seattle and from Minneapolis. And you can see that on the edges [of the distribution], there is a lot of differences [between the groups], but there is very little quantity of samples that are outside the major mid part of the graph, where [the simulated samples] mostly is located. And since the middle is zero, which is 'no difference', a lot of the data is concentrated around 'there is no difference' … It seems there is very little sample data showing a lot of differences between the two airports. So I would, without looking at the numbers, it suggests to me that my null hypothesis is correct. (Appendix P02-B2, 03:57 – 05:08).

Additionally, Tal did not seem to focus on the *p*-value while monitoring the progress of the comparison between groups, nor while interpreting results or justifying the interpretation that there is not much difference between the groups. For example, while completing the VSE problem, Tal stated that "We have the *p*-value which is very low, which corroborates the null hypothesis that there is no difference in the means. I think the difference was like 3 point something percent, which is probably not very significant" (Appendix P02-B1, 09:00 – 09:17). In the *video-cued interview*, Tal explained that they

saw the *p*-value as a number without context (Appendix P02-C, 15:15 – 16:35), and thus not as important as effect sizes, confidence intervals, or group means in monitoring nor evaluating. Thus, it appears that Tal thought of significance not in terms of statistical significance but rather in terms of a practical or meaningful significance based on the context of the problem.

While Tal did occasionally monitor and think of *p*-values, as was the case in the *statistical testing task,* while completing the *statistical testing interview,* Tal seemed not to think about *p*-values or include them in the monitoring of the test. Instead, the driving factor in Tal's monitoring seemed to be the context of the problem and the theory or prior knowledge underlying that context. An example of this thinking was evident when Tal attempted to answer a question about the difference in PISA scores between Finnish students and Spanish students (see Appendix H2 for the *statistical testing interview* stimuli). Tal explained that, "knowing what I know of Finnish and Spanish, I would say yes, because everybody says great things about Finnish education" (Appendix P02-C, 15:45 – 15:55). Thus, it appears that Tal's prior knowledge that "everybody says great things about Finnish education" was used by Tal to monitor whether the statistical results may have been generated appropriately.

4.4.1.3 Tal's Evaluating of Statistical Tests

Tal's notion of "sensible results" appeared to dominate Tal's evaluating. However, when Tal's evaluating did result in adjustments to Tal's plan, the adjustments did not seem to result in changes to computations Tal would make, but rather resulted in the re-examination of the data to ensure integrity, and subsequently the re-evaluation of a theory or research question, with Tal questioning its "sensibility". As previously mentioned, after

generating a *p*-value, Tal focused on the context at hand to determine practical significance. As Tal described, "You cannot look at the *p*-value without … a number by itself doesn't mean [anything], you have to have context. One of the first things that I do is, especially when you have a difference in averages, I'm going to see if that means if it's really different" (Appendix P02-C, 15:15 – 16:35). Tal's process for determining whether a difference is 'really' different seemed not to rest on *p*-values or confidence intervals, nor any statistical procedure, but rather on the underlying context.

More generally, while completing the *concept mapping task*, Tal provided an insight into the way they thought about evaluating statistical tests. As previously described, if the results were not sensible, then Tal suggested that they would go back to the theory, tweaking it, or perhaps tweaking the research question, until they found results that made sense given a prevailing theory.

4.4.2 Tal's Thinking about Null Models

While Tal's thinking focused on the use of descriptive statistics, it seemed that Tal did make considerations for sampling variability, although not in the form of a null model. It seemed that Tal ascribed uncertainty to the possible hypothetical parameters given the observed statistics, rather than ascribing uncertainty to the possible statistics given a hypothesized parameter (the norm in NHST). Therefore, it is worth considering how Tal thought about sampling variability and what Tal believed the randomization dot plot based on a null hypothesis represented.

While completing the Airplane Delays problem, Tal appeared to confuse the randomization dot plot with a bootstrap dot plot – the key difference is that a bootstrap dot plot is centered on the observed sample statistic, while the randomization dot plot is

centered on the null hypothesis parameter. However, Tal's thinking about the dot plot was otherwise consistent with thinking about sampling variability through the examination of the bootstrap dot plot. Tal focused on the entire distribution, explicitly acknowledging its shape, center, and spread, the three characteristics of a distribution as instructed in EPSY 5261 – this is seen not only in the audio recording, but also in the gaze recording of Tal's eyes, showing clearly that Tal was looking at the entire distribution (Appendix P02-B2, 03:57 – 05:08). Tal also focused on the shape, center, and spread of dot plots presented in the *statistical testing interview*. For example, in answering the question "Is average commute time in Atlanta and St Louis the same?", Tal commented that "… we have a normal distribution, uhh, and pretty much every data is like between -2 and 2" (Appendix P02-C, 01:01 – 01:14). Similarly, when answering the question "Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?", Tal stated that, "… we have a normal distribution, 97% of the information is between 27 and 29 …" (Appendix P02-C, 03:27 – 03:33).

Additionally, the way Tal described their thinking while examining the randomization dot plot was further evidence that Tal was thinking about sampling variability. As Tal described, "… you can see that on the edges, there is a lot of differences, but there is very little quantity of samples that are outside the major mid part of the graph, where [the simulated samples] mostly is located. And since the middle is zero, which is 'no difference', a lot of the data is concentrated around 'there is no difference' …" (appendix P02-B2, 03:57 – 05:08). Thus, it appeared that Tal, in thinking about the entire sampling distribution, was inherently thinking about the variability of that sampling

distribution, and thus explicitly acknowledging the role of sampling variability in statistical testing, despite misconstruing the randomization dot plot to be a bootstrap dot plot.

However, this thinking appeared to be local to the simulation-based output Tal interacted with. When examining output from R in the *statistical testing task* and the *statistical testing interview*, there was no hint that Tal was thinking about sampling variability. For example, after interpreting the summary statistics generated by the favstats() function in R, Tal stated, "Now, to make sure that what I implied is correct, I'll have to run the test" (Appendix P02-B1, 05:58 – 06:09). Why did Tal feel that it was important to 'make sure', and what did Tal think running the test would achieve? Tal never actually commented upon this in either the *statistical testing task* or the *video-cued interview*.

However, Tal's interpretation of the results in the VSE problem may provide a clue as to Tal's thinking. There, Tal stated that "We have the *p*-value which is very low, which corroborates the null hypothesis that there is no difference in the means. I think the difference was like 3 point something percent, which is probably not very significant" (Appendix P02-B1, 09:00 – 09:17). Tal commented upon and interpreted the *p*-value, but seemingly returned to considering the effect size and the context of the problem when drawing a conclusion about the "significance" of the results. First, it seemed that Tal was using the term "significant" to mean something like "important" or "meaningful". Additionally, Tal's preference for thinking about the meaningfulness of results, and doing so through thinking about effect sizes, suggests that perhaps Tal considered *p*-values a mere formality. This might also explain why Tal commented in the *video-cued interview* that "a

number by itself doesn't mean [anything]" (Appendix P02-C, 15:15 – 16:35) when commenting upon interpreting *p*-values.

Perhaps the most that can be said is that how Tal determined what is and is not sensible relied on an implicit acknowledgement of sampling variability in relation to the specific predictions of a research question, but not formally. Yet, it is striking that Tal consistently and thoroughly thought about the bootstrap dot plot when examining results generated by StatKey, but did not exhibit analogous thinking when examining results generated by R.

4.5 Participant Five – Ade

Ade took EPSY 5261 in the Fall of 2021, seven months before participating in the study. Ade did not take any other statistics course between completing EPSY 5261 and participating in this study. However, Ade did take a statistics course while an undergraduate student which was very similar in nature to EPSY 5261. Furthermore, Ade took a research methods course in the Spring of 2022, which did have a small unit on statistical methods, but this was limited to the examination of means, graphs, and *p*-values from *t*-tests. Aside from homework assignments in EPSY 5261 and one assignment in that research methods course, Ade had never done their own statistical analysis for any of their own projects.

Ade commented that despite perceiving that statistics did have value, Ade personally did not feel a need to do statistics. Ade did not plan on ever doing their own statistical analysis, instead only finding it important to be able to interpret results from statistical analyses. Despite this, Ade admitted to usually skipping the methods and results

sections of papers, instead jumping straight to the conclusion, in which Ade expected to be told whether there was a 'significant finding'.

Ade also admitted to having different study habits for EPSY 5261 than for their other classes. In EPSY 5261, Ade studied just before the exams, whereas for other courses, Ade's studying was more spread out and involved a deeper dive into the material, due to a greater interest in the content. Ade commented that for EPSY 5261, they typically put in only an hour or so of studying before the exam, and then promptly forget about the content. Important to remember is that Ade was instructed not to prepare or review any materials prior to participating in this study.

4.5.1 Ade's Thinking When Conducting Statistical Tests

To Ade, the purpose of a statistical test was to produce a *p*-value and confidence interval, which could then be used to answer a research question (see Figure 17 and Appendix P05-A, 01:09 – 01:20). Also important in statistical tests were null and alternate hypotheses. Ade saw randomization tests and *t*-tests as two methods of producing *p*-values, randomization tests doing so through resampling, and *t*-tests doing so through equations. Finally, Ade noted that "the lower the *p*-value the better" (Appendix P05-A, 01:14 – 01:15). In thinking about *p*-values and confidence intervals, Ade drew conclusions about whether or not the null hypothesis was supported. However, at one point Ade also stated that "if we're just doing difference in means, I feel like I just take these [the estimate of the mean for each group] then" (Appendix P05-B1, 05:15 – 05:26). Thus, it is unclear when Ade thought a statistical test would indeed be necessary.

4.5.1.1 Ade's Planning of Statistical Tests

Ade's plan for statistical testing began with generating summary statistics. In the *statistical testing task*, after reading the prompt for the VSE problem and identifying that "we're doing difference of means" (Appendix P05-B1, 00:46 – 00:49), Ade determined that, "well for each group, I'm assuming we're going to have to get the relevant statistics where it shows means, standard deviations, all of that" (Appendix P05-B1, 01:56 – 02:03). Ade then created a histogram, although at first, Ade was unsure what type of graph to make, commenting that, "I'm assuming we'll have to do a graph … I don't know which one" (Appendix P05-B1, 00:53 – 00:56).

Next, Ade identified that "this is when testing hypothesis comes into play. Does a difference exist between the two?" (Appendix P05-B1, 09:11 – 09:20). Ade first specified the null and alternative hypothesis, stating, "so I know the [null] hypothesis would be like, 'cs is equal to engineering' and that would be like the hypothesis is false [based on the difference in means], but 'cs does not equal engineering', then the [alternative] hypothesis is supported." (Appendix P05-B1, 09:41 – 10:00).

After generating output, Ade focused on the 95% confidence interval estimate, and also inspected the *p*-value, before stating, "there's evidence to support that the [alternative] hypothesis is true and there is a difference in mean salaries between the two groups" (Appendix P05-B1, 11:15 – 11:43), because "0 is not included within the confidence interval" (Appendix P05-B1, 12:55 – 12:59).

Thus, Ade's plan for conducting a statistical test seemed to be to (1) generate summary statistics, (2) generate histograms, (3) specify a null and alternate hypothesis, (4) generate a confidence interval and *p*-value, and (5) compare the null hypothesis parameter to the confidence interval.

4.5.1.2 Ade's Monitoring of Statistical Tests

Ade admitted to having difficulty remembering how to monitor the process of conducting statistical tests, commenting at one point that, "as far as testing that to figure out if there is a difference that exists, I do not remember this part" (Appendix P05-B1, 10:04 – 10:10). However, Ade's monitoring was somewhat evident as computing summary statistics and generating histograms led up to Ade running a *t*-test.

Having generated summary statistics for each group, Ade stated that "engineering does have a higher average salary, but there is variability, like different variability" (Appendix P05-B1, 06:12 – 06:27). Initially, Ade seemed to have planned only to compute means for each group to answer the research question, but seeing the standard deviations prompted a re-evaluation of this plan (Appendix P05-B1, 05:46 – 06:02).

Similarly, having generated histograms for each group, Ade stated that "engineering just from the graph does look more … they're both like normal distributions, at least I think so" (Appendix P05-B1, 07:43 – 07:54). From the gaze recording and screen recording of this clip, we see that in saying "look[s] more", Ade was comparing the location of the mode of each distribution. Ade subsequently visually inspected the tails of the distribution, which seemed to prompt a re-evaluation of Ade's plan.

Additionally, when conducting the statistical test, Ade monitored the *p*-value in the *t*-test and randomization test. Specifically, Ade utilized the *p*-value to monitor the level of confidence they had in the test results. For example, in the *statistical testing interview*, Ade at one point commented that, "well it has a low *p*-value so I feel like this is a good sample. And then, just the means alone, 81 and 67 are pretty different" (Appendix P05-C, 17:36 – 17:46). It seemed that Ade, in saying the "sample" was "good", was thinking about the

credibility of the inference being drawn by comparing the sample means directly. However, this comment may also be a result of Ade thinking about the desirability of low *p*-values in relation to conventional declarations of statistical significance.

More generally, Ade seemed to be unsure in monitoring their plan for statistical testing. For example, in the *statistical testing task*, Ade commented that, "I feel like I should be looking at standard error for some reason … because it would be the mean plus or minus the standard error, and we're seeing if a difference exists" (Appendix P05-B2, 05:02 – 05:40). Here, it seems that Ade perhaps initially had not planned on thinking about the standard error, but in the middle of the task remembered the connection between standard error and the process for constructing a confidence interval estimate, which thus led to Ade use the standard error to monitor their plan for conducting the statistical test. However, it is unclear whether any specific statistical output Ade was thinking about prompted this recollection, or whether it was simply Ade remembering additional statistical content as the tasks progressed.

Thus, Ade's monitoring of a statistical test for a difference in means seemed to hinge on the thinking about the sample distributions and any differences between them as described by summary statistics and data visualizations, as well as the *p*-value.

4.5.1.3 Ade's Evaluating of Statistical Tests

Ade's evaluation of their plan for conducting a statistical test stemmed from the same components to Ade's monitoring – variability expressed by the sample standard deviation and the range in the histogram, as well as the *p*-value. Ade's monitoring of the sample standard deviation prompted an evaluation of Ade's plan, as Ade explained in their own words, "engineering does have a higher average salary, but there is variability, like

different variability, like the standard deviations, this one [the standard deviation for salaries for the computer science group] is higher. I think I should graph them" (Appendix P05-B1, 05:46 – 06:27). Thus, the differences in sample standard deviations prompted Ade to change their initial plan, and proceed with generating histograms.

Similarly, having seen the variability in the histograms for each graph, Ade again re-evaluated their plan, stating "I think this is when testing hypothesis comes into play. Does a difference exist between the two?" (Appendix P05-B1, 09:11 – 09:20). Thus, the differences in the sample distributions prompted Ade to change their initial plan, and proceed with conducting a statistical test.

4.5.2 Ade's Thinking about Null Models

Ade connected null models and the randomization dot plot to the null hypothesis (although not at first), but was not able to fully recollect the role null models played in testing. Initially, Ade focused on the normality of the randomization dot plot. For example, in the *statistical testing task*, Ade commented that "I feel like I need to generate enough samples so it is a normal distribution" (Appendix P05-B2, 03:04 – 03:11). Similarly, during the *statistical testing interview*, Ade described one randomization dot plot as "a very normal distribution" (Appendix P05-C, 08:11 – 08:14). However, it is unclear why Ade thought it was important for the randomization dot plot to be normal.

Partially through the *statistical testing interview*, Ade recollected that the randomization dot plot was based on the null hypothesis. This occurred while Ade was answering the question "Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?". After first interpreting the middle 95% of the randomization dot plot as if it were a confidence interval, Ade stated, "so I feel like, something about the null being 28.6 that,

when you do a randomization test, it is going to center around that always" (Appendix P05-C, 05:22 – 05:34). Ade then reasoned that instead of determining whether the null hypothesis parameter lies within this middle 95% of the randomization dot plot, "I think we're looking at like, maybe how far from the [hypothesized] mean [the confidence interval] strays" (Appendix P05-C, 05:51 – 05:59).

Having connected the null hypothesis to the center of the randomization dot plot, there are two lingering questions in terms of how Ade thought about null models: (1) How did Ade think about the standard error? And, (2) What did Ade think the *p*-value represented?

4.5.2.1 Ade's Thinking about Standard Error

Ade appeared to think about the standard error as a quantification of uncertainty related to inferences. Procedurally, Ade's thinking about standard error was tied to its role in computing a confidence interval. For example, in the *statistical testing task,* Ade commented, "I feel like I should be looking at standard error for some reason … because it [the confidence interval] would be the mean plus or minus the standard error, and we're seeing if a difference exists, so if it equals 0 then, like if that parameter [recta, the confidence interval] has 0 in it, then the null hypothesis is … mmm … I'm trying to think of like what the conclusion would be … that like there's not sufficient evidence that a difference exists between the two means then" (Appendix P05-B2, 05:02 – 05:40).

In the *statistical testing interview,* Ade further commented about the specific role the standard error had in quantifying precision, through its role as part of the confidence interval. Ade stated, "I feel like there's more room for error [when the CI is far apart from the mean]" (Appendix P05-C, 06:38 – 06:43). However, Ade did not seem to exactly

remember how the standard error and confidence interval quantified variability, commenting that "I think it [the confidence interval] takes into account outliers, so we're using more of the center of the data … and like standard deviation, that's part of this" (Appendix P05-B1, 13:25 – 13:42). Thus, it did seem that Ade thought about the role the standard error played in quantifying uncertainty, through its specification of the width of the confidence interval.

4.5.2.2 Ade's Thinking about P-values

Ade seemed to primarily think about *p*-values in terms of their link to confidence levels. In the *video-cued interview*, Ade explicitly noted this equivalence, writing "$p < .05$ = 95% Confidence Interval" (Appendix P05-D3). To Ade, a *p*-value of .05 "would take off .25 and .25 [sic] percent of the data [from the tails of the sampling distribution]" (Appendix P05-A, 02:30 – 02:39), achieving "95% confidence that the datapoint [sic] will fall within the interval" (Appendix P05-A, 02:46 – 02:53). Ade further commented on this relationship in the *statistical testing interview*, stating that, "I think that [the *p*-value] has something to do with error, so a lower *p*-value is better, because at .05 then you're at 95% confidence" (Appendix P05-C, 02:29 – 02:47).

Thus, it seemed that Ade thought of the *p*-value as the complement of the confidence level. When answering the question "Is the average number of body piercings UMN undergrads have equal to 2?" during the *statistical testing interview* (see Appendix H2 for the *statistical testing interview* stimuli), Ade commented that, "when I did this for my class, I feel like we were given smaller *p*-values, and this is a pretty big *p*-value, so I guess because 2 is included within this interval you could say that the null is supported, but

with 70% confidence, so it's not really that big of a deal." (Appendix P05-C, 16:03 – 16:34).

Despite thinking about the *p*-value in terms of the complement of the confidence level, Ade appeared to think about *p*-values as contributing distinct information from the confidence level. Specifically, Ade appeared to prefer small *p*-values as an indication of the accuracy or validity of the test. For example, in the *statistical testing interview*, Ade commented that, "well there's a low *p*-value so I feel like this is a good sample" (Appendix P05-C, 17:36 – 17:40). After inspecting the *p*-value to determine whether or not the test is a 'good' test, Ade then compared the null hypothesis mean to the confidence interval, to determine whether the parameter was contained within the interval.

4.6 Participant Six – Aan

Aan took EPSY 5261 in the Fall of 2021, seven months before participating in this study. Aan did not take any other statistics course after EPSY 5261. Aan also had not done any statistical analysis since completing the class, but had studied quantitative analysis in another class a year before completing EPSY 5261. Furthermore, quantitative work was not heavily featured in Aan's field of study, as Aan's field and Aan's own research was mainly qualitative. When quantitative analysis was included, it was almost always in the form of either descriptive summary statistics or a *t*-test comparing two groups. Aan's knowledge of statistical software was also limited to software learned in EPSY 5261. Aan admitted forgetting a lot about R since the end of EPSY 5261. As with other participants, Aan was asked not to prepare prior to participating in this study.

4.6.1 Aan's Thinking When Conducting Statistical Tests

To Aan, the purpose of a statistical test was to compare a hypothesis to some observed evidence in the form of data that is collected. However, Aan's thinking in this study almost entirely focused on point estimates of means. In the *video-cued interview*, Aan stated that, "I just remember a few key words, like the mu, or the standard deviation, standard error, so that's it, but I don't know what [they are]. I forgot about it! Even for *p*-values, I know the parameter .05, equal, above, or less, but I forgot what that means. But the only one thing [I know is] the mean, I can determine that the mean is different, or the mean is not different" (Appendix P01-D, 05:10 – 05:40). During the *statistical testing task*, there were many moments where Aan went silent, and Aan's eye gaze patterns suggested that Aan seemed to be systematically scanning the entire screen, perhaps attempting to recollect important pieces of information and how they might relate to the task at hand.

4.6.1.1 Aan's Planning of Statistical Tests

Aan's thinking in terms of a general plan for conducting statistical tests is evident in the concept map that Aan created (see Figure 18). As seen in Aan's concept map, Aan's plan for a statistical test began with a question and two competing theories. Next, Aan's plan specified that data should be collected. The two competing theories then led Aan to specify a null and alternate hypothesis. However, Aan's null hypothesis was a "not equal to" hypothesis and while Aan's alternate hypothesis was an "equal to" hypothesis (Appendix P01-A, 00:27 – 00:38).

After the specification of the null and alternate hypotheses, Aan proceeded to compute sample means for each group. In the *statistical testing task*, after reading the research question for the VSE problem, Aan succinctly laid out a purpose and associated plan for answering the question: "We need to compare if [the average salary in each group]

is the same or different. I think I should compute the mean for each group first, and then run a *t*-test" (Appendix P01-B1, 01:56 – 02:20).

However, despite the clear articulation of competing hypotheses in Aan's concept map, during the *statistical testing task,* Aan did not specify or articulate hypotheses, nor did Aan comment upon what those hypotheses may be in the *video-cued interview*. Instead, Aan's emphasis on the descriptive statistics suggested that hypotheses did not play a critical role in Aan's thinking while conducting statistical tests in the *statistical testing task*. Rather, it appeared that Aan's focus on generating and interpreting sample estimates.

Although Aan's concept map and Aan's stated plan both included generating a *p*-value from a *t*-test, in the *statistical testing task,* Aan was ready to make a conclusion for the VSE problem after only examining the means (and standard deviations) of each group: "I think there is a difference in the average salary for the computer science and engineering [majors], but the difference is not too huge, just a tiny difference" (Appendix P01-B1, 10:52 – 11:11). Only after being reminded by the researcher that Aan's original plan was to compute the means and then run a *t*-test did Aan proceed to conducting the statistical test. Finally, Aan's plan ended with the extraction of a *p*-value from the *t*-test, using the *p*-value to make a decision about which hypothesis is correct (Appendix P01-A, 01:50 – 02:05).

4.6.1.2 Aan's Monitoring of Statistical Tests

Aan's monitoring of statistical tests appeared to primarily consist of a comparison of the conclusion drawn from comparing two means to each other (either the null hypothesis and the single group sample mean, or the means from two groups) with the conclusion drawn from the interpretation of a *p*-value. This is most clearly evident in Aan's

comments during the *video-cued interview*. There, Aan stated that "in the *t*-test, I think I'm looking for the *p*-value … this *p*-value is less than .05" (Appendix P01-D, 08:06 –08:20). At this point the researcher prompted that, "but it looks like you were looking at the means first", to which Aan responded "Yes, because, I can compare the means very directly, but the *p*-value, I told you, I'm [confused] about the three … is it above or less or equal [to .05], what does it mean, behind this code? Of course, I'm just thinking about, the only thing I'm very confident that I know is the mean. The mean very directly shows the result. They're equal, they're different. And then I'm thinking about the *p*-value" (Appendix P01-D, 08:25 – 08:58).

While Aan perhaps wanted to rely on *p*-values to monitor the statistical test, due to uncertainty with regards to the proper interpretation of *p*-values, Aan appeared to fall back on interpreting the means instead. This is also reflected in the gaze recording of Aan's eye movements during the *statistical testing task*, in which Aan spent more time looking at the means of each group than the *p*-value or the 95% confidence interval estimate when interpreting results from the *t*-test in R (see Figure 19 for a heat map summary of the locations on the screen Aan looked at the most, Figure 20 for a clearer image of the R output Aan was looking at, and Appendix P01-B1, 13:20 – 14:30, for the gaze-recording of this moment). Thus, despite Aan predominantly focusing on the means of each group, it might still be the case that Aan's thinking, and internal logic of statistical testing, did indeed include interpreting *p*-values, especially as a tool for monitoring.

4.6.1.3 Aan's Evaluating of Statistical Tests

Aan's evaluating of their plan for the statistical test was almost entirely absent from the traces of Aan's thinking captured by this study. Throughout the *statistical testing task*

and the *statistical testing interview*, there appeared to be little changes to Aan's original plans. This may have been a result of the fact that Aan admitted that they had forgotten quite a lot of the information they were taught in EPSY 5261, leading to a somewhat simple plan. Indeed, Aan's plan, as operationalized by Aan in the *statistical testing task*, almost entirely consisted of computing sample means for each group, and nothing more. In the *statistical testing task*, it was the researcher who prompted Aan to conduct the *t*-test, and not Aan's own evaluation or adherence to Aan's original plan. Thus, it is unclear if, and if so how, Aan evaluated their thinking while conducting statistical tests or interpreting results from statistical tests.

4.6.2 Aan's Thinking about Null Models

The preponderance of evidence indicates that Aan did not consider nor likely had a concept of a null model that they were thinking about while conducting statistical tests nor while interpreting results from statistical tests. The most consistent thing that Aan appeared to do was to think about the sample mean or the difference between sample means. Even when Aan referenced the mean of the randomization dot plot, Aan's statements seem to indicate that Aan believed the randomization dot plot was estimating what is true in the real world based on the sample data. In addition, Aan typically made absolute judgments without any consideration of sampling variability.

4.6.2.1 Aan's Thinking about Sampling Variability

While conducting a statistical test in R in the *statistical testing task*, Aan did comment on the sample standard deviations while examining summary statistics, stating "oh [the standard deviations] is a lot" (Appendix P01-B1, 09:52 – 09:55). However, it did not appear that Aan factored the sample standard deviations into any further analysis. Aan

explicitly stated as much before drawing a conclusion – "Back to the research question … because the question is talking about the average salary, so back to the data, this should be reflected in the means" (Appendix P01-B1, 11:16 – 11:24).

However, Aan's interpretation of the results from the *t*-test in the while completing the VSE problem might suggest that Aan was thinking about sampling variability. In examining the output of the t_test() function in R, Aan stated, "Okay, sample estimates … here is the result. It only calculated the mean. So, …, they're different!" (Appendix P01-B1, 13:23 – 13:36). However, once Aan saw the *p*-value, Aan began to hedge: "… and the *p*-value is .04227, let me think, I think this *p*-value is too small, it's small … There is not a big difference between the two groups" (Appendix P01-B1, 13:40 – 14:20). Was this due to a consideration of sampling variability?

Aan's comments in the *video-cued interview* provide some insight to this question. There, Aan stated that "The mean very directly shows the result. They're equal, they're different. And then I'm thinking about the *p*-value. And the last one I think [I'm still] looking for is the standard error. Because according to [ESPY 5261], there are only three things we can determine, one thing is the mean, *p*-value, and the sd [standard deviation]" (Appendix P01-D, 08:50 – 09:20). While Aan did mention the standard deviation, at no point in the actual test did Aan comment upon the standard deviation or the standard error. Furthermore, Aan's description that the mean showed the result directly as either "equal" or "different", and Aan's subsequent examination of the *p*-value, suggests that Aan may have been using the *p*-value to determine the magnitude of the difference between groups. As seen above, Aan claimed that the means were different, but then stated that there was not a big difference after examining the *p*-value. If Aan was explicitly considering

sampling variability, there did not appear to be a clear trace of this thinking in the data collected.

However, it should be noted that it is possible that in determining that "computer science and engineering is different", Aan may have been thinking about the *p*-value, while when determining that "there's not too much of a big difference", Aan may have been thinking about the effect size. Either way, whether Aan's statement about "not too much of a big difference" was based on the *p*-value or the effect size, Aan did not explicitly comment on a quantification of variability, and did not seem to be explicitly thinking about sampling variability.

4.6.2.1 Aan's Thinking about the Randomization Dot Plot

Aan did not seem to remember how to think about simulation-based statistical testing, and did not seem to remember how to think about the randomization dot plot. In the second part of the *statistical testing task* in which Aan utilized a simulation-based approach to completing the Airplane Delays problem, after spending a few minutes seemingly in an attempt to recollect what to do, Aan whispered, "I'm thinking [about] which parameter can represent this null hypothesis, [which] is correct. I think this time it's not the mean, maybe the standard error … Okay let's do the standard error. [There is] no difference. So, then the null hypothesis is correct" (Appendix P01-B2, 05:49 – 06:24). Aan appeared to see that in the top-right corner of the randomization dot plot, StatKey provided three numbers: "samples = 5000", the number of simulated trials; "mean = -0.0031", the mean of all of the simulated trials, which will always be very close to the null hypothesis mean value; and, "std. error = 0.357", the standard error of the sampling distribution under

the null hypothesis, computed as the standard deviation of the simulated trials' means (see Figure 21).

Why did Aan state that "I think this time it's not the mean"? It is unclear, especially since Aan had previously stated "we will need to compute the average delay time for each group, one group is for Minneapolis, another group is for Seattle. Hmm, okay, I think that's the same thing [as the VSE problem]" (Appendix P01-B2, 01:24 – 01:45). However, later in the task, Aan stated that "because the two simulated groups' average is very tiny, it's almost equal to 0, so I think there is no difference between those two groups. They are almost the same" (Appendix P01-B2, 04:34 – 04:55).

However, the fact the Aan noticed the standard error is not surprising, as Aan commented in the *video-cued interview* that "… according to [ESPY 5261], there are only three things we can determine, one thing is the mean, *p*-value, and the sd [standard deviation]" (Appendix P01-D, 09:10 – 09:20). Yet, Aan's thinking about the standard error is unclear. Was Aan making an inference based on the magnitude of the standard error, concluding that "the null hypothesis is correct" because a standard error of .357 suggested so? It is unclear, and perhaps all that can be said is that Aan knew that the standard error was important, but did not remember how to think about it.

In the *statistical testing interview*, specifically when interpreting output from StatKey, Aan similarly focused on thinking about the "mean" and "standard error", again displaying the same thinking Aan demonstrated in the *statistical testing task*. For example, when examining results for the question "Is average commute time in Atlanta and St Louis the same?" (see Appendix H2 for the *statistical testing interview* stimuli), Aan stated, "This shape [the randomization dot plot] is a bell shape. Soo, I think also the mean is very tiny,

but the standard error … is big?! So I think … they are not the saaaaame?" (Appendix P01-C, 01:50 – 02:15). Aan's response indicated that Aan was perhaps uncertain about their thinking, but nevertheless, Aan still focused on the mean and standard error.

Aan's thinking about the mean and standard error in interpreting output from StatKey was similarly evident in the gaze recordings throughout the *statistical testing interview*. For example, when answering the question "Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?", Aan predominantly looked at the mean and standard error (see Figure 22).

What did Aan believe the randomization dot plot represents? Perhaps the answer best supported from the empirical trace of Aan's thinking is simply that Aan did not remember what the randomization dot plot represents, nor where it came from, nor its role in statistical testing.

Chapter 5: Discussion

How do graduate students who have completed a master's level introductory statistics course utilizing a simulation-based inference (SBI) curriculum think when they conduct statistical tests? That question was the genesis of this study. More specifically, the two research questions that motivated the task design, data collection methods, and data analyses for this study were:

(1) What is the nature of graduate students' thinking when conducting statistical tests?

(2) Do graduate students think about null models when conducting statistical tests, and if so, how?

Six graduate students were recruited seven months after they completed such a course to participate in a series of tasks designed to elicit multiple aspects of their thinking. First, these six students all created a concept map for the logic of a statistical test. The audio recordings as well as the concept maps each student drew were collected and analyzed. The six students then were presented with two problems and asked to solve one utilizing traditional parametric-based methods in R and the other utilizing simulation-based methods in StatKey. The audio recordings, screen recordings, and eye gaze recordings from this task were collected and analyzed. Next, the students were presented with output from a series of ten statistical tests – five generated by R and five by StatKey – and were asked to interpret the results. Once more, the audio recordings, screen recordings, and eye gaze recordings from this task were collected and analyzed. Finally, the students watched the screen and eye gaze recordings from when they used R and StatKey to conduct analyses to solve each of the two problems presented earlier, adding additional detail to retrospectively explain their thinking, as well as discussing and verifying the inferences made by the researcher. Here, audio recordings and screen recordings were collected and analyzed.

These data artifacts were analyzed utilizing an interpretivist epistemological stance within a multiple descriptive case study approach. Thinking was a priori defined as metacognitive self-regulation, operationalized through the component actions of planning, monitoring, and evaluating. However, the focus of the analysis was on using this framework to credibly explain the phenomenon of statistical thinking as experienced by each student. To achieve this, data artifacts were examined through a constant comparative process with inductive coding, with one data artifact examined at a time, before moving

through all artifacts for a task, and then all tasks for a single student, before once more reviewing all data to faithfully describe each students' thinking.

5.1 Research Question 1: What is the nature of graduate students' thinking when conducting statistical tests?

Each of the six students who participated in this study seemed to think somewhat differently when conducting statistical tests. These differences were apparent even in their perceived purpose of conducting a test.

Two participants, Ade and Chau, saw the purpose of a test as being the determination of whether or not a relationship or difference exists between two groups or two factors, which could answer specific research questions. Tal saw the purpose of a statistical test as being to confirm a theory, through an iterative process, until a theory is plausible given some observed data. Jaci similarly saw a test's purpose as being about theories, but as opposed to Tal, Jaci saw the purpose of a statistical test as the determination of whether the null hypothesis or the alternative hypothesis is true. Jaci also provided another explanation, one based on whether the observed difference was "naturally occurring", a term Jaci used to imply contextual consistency based on a small margin of error. Aan saw the purpose of a statistical test as the comparison of evidence to a hypothesis, but did not articulate or think about sampling variability as part of the process. Finally, Kei seemed not to fully or clearly understand a purpose of a statistical test, commenting that one could simply "look at" the data and just tell whether or not there was a difference.

As each participant saw a different purpose for statistical tests, the rest of their thinking also starkly differed from each other, in terms of their planning (i.e., goal setting and the selection of appropriate strategies), monitoring (i.e., self-checking on the progress

of actions towards achieving one's goals), and evaluating (i.e., revising one's goals and strategies when necessary).

5.1.1 The Nature of Graduate Students' Planning When Conducting Statistical Tests

Despite all six participants' planning when conducting statistical tests being different, their plans all shared some common features. For example, nearly all participants' plans led to the computation of a confidence interval or *p*-value. However, there are few similarities beyond this.

Aan's plan was very similar to a purely Neyman-ian likelihood ratio-based decision theoretic model. Aan's plan began by specifying two hypotheses. Then, Aan collected and examined data, computing means for each group. Aan then quickly proceeded to generating an estimated difference between each group. This estimate was then compared to the two candidate hypotheses, with one being retained as the best hypothesis. Interesting to note is that Aan's plan did not explicitly account for any uncertainty through the specification of a probability distribution.

Tal's plan was very similar to the type of thinking consistent with the likelihood school of statistics, or perhaps even confirmation by instances – Tal planned to examine data and determine whether some given evidence supported a particular hypothesis. Tal's plan began with the specification of a theory, or rather, a research question that is laden with a general theory but contained a specific and contextualized claim. Tal's plan then led to the collection and analysis of data, computing means for each group and most importantly, a confidence interval estimate. Tal then considered the "sensibility" of the results. One interesting note is that Tal's determination of sensibility of the results might be construed as consistent with the hypothetico-deductive approach of the classical school

– a theory specifies which observed data might be construed as sensible through a null model. However, Tal's determination of sensibility seemed not to be in terms of any observed data, but rather the sensibility of a research question. Therefore, Tal's thinking was more consistent with confirmation by instances, which considers whether evidence supports the development of a hypothesis. However, Tal's planning seemed to clearly depend on an explicit specification of a probability distribution. As opposed to the form of a null model, Tal's plan for a statistical test rested upon a bootstrap distribution, which Tal thought about in a manner akin to thinking about a likelihood function and from which the confidence interval is extracted.

Kei's plan for a statistical test was highly driven by the set of procedures Kei was taught in EPSY 5261. Kei's plan began with the data and examining each group separately. Kei first computed sample statistics and generated a graph for one group, ideally a histogram, to contextualize the data and develop a real-world understanding. After doing the same for the second group, Kei then computed the estimated difference between groups. Kei's plan did not seem to automatically progress beyond this point. However, in contexts in which Kei might be expected to produce inferential statistics, Kei's plan then led to the computation of either a confidence interval or a *p*-value. However, Kei utilized these statistics as supplementary to the estimated difference between groups in determining whether there was a real-world difference between groups. Thus, Kei's plan, like Aan's, did not explicitly account for any uncertainty through the specification of a probability distribution.

Jaci's plan for a statistical test was somewhat similar to the type of thinking utilized in the likelihood approach to statistics and in confirmation by instances, although as

opposed to Tal, Jaci's plan did still contain elements from the classical approach to statistical testing, particularly decision-theoretic elements in selecting between null and alternate hypotheses. Jaci's plan for a statistical test began with data. Next, on the basis of the data and context, Jaci determined which statistical test to conduct. Jaci then wrote the null hypothesis, before generating a simulated sampling distribution. However, as opposed to a null model, Jaci's plan called for a simulated sampling distribution based on the data, i.e., a bootstrap dot plot, which Jaci interpreted and interacted with in a manner similar to the way in which one might interact with a likelihood function. From this, Jaci extracted a confidence interval (or alternately obtained a *p*-value automatically through statistical software) and determined whether the observed difference was a significant one. Here, Jaci's plan was not to determine statistical significance, but rather a form of practical significance, based on the width of the confidence interval as a measure of the data's precision in supporting a particular hypothesis (something akin to the degree to which evidence contributes to the development of a hypothesis in confirmation by instances).

Ade's plan for a statistical test, like Aan's and Kei's, did not explicitly specify a probability distribution, although Ade did connect the null hypothesis to the center of the randomization dot plot in StatKey and thus explicitly acknowledged the null model. Ade's plan began with computing summary statistics for each group, before also creating graphs for each group. Ade then specified a null hypothesis, before obtaining a *p*-value and/or confidence interval. Based on the *p*-value, Ade determined whether the null hypothesis should be rejected. Thus, while consistent with the classical approach to statistical testing, Ade's plan was very product-oriented. That is, Ade's plan seemed to focus on the

extraction of a *p*-value, without any explicit steps specifically related to null models, nor any explicit thinking about the process by which a *p*-value is computed.

Like Ade, Chau took a very product-oriented approach to statistical testing, albeit with a much more involved plan than Ade's plan. Chau's plan began with a research question, which specified the null and alternate hypotheses as well as the significance level against which the *p*-value would be interpreted. Chau also saw these research questions as fully specifying the type of test that would be conducted. Chau then examined the sampling strategy employed in the data generating process to ensure that it was a random sampling strategy, as a check on the assumptions of the *t*-test. Next, Chau computed summary statistics, seemingly to get a sense of the data. However, Chau also generated histograms, but mainly to check the normality assumptions of the *t*-test (when needed, based on sample size). Chau's plan then led to conducting a statistical test and the extraction of a *p*-value. This *p*-value was then compared to the pre-specified significance level to determine whether the null hypothesis should be rejected. Thus, while a faithful application of the NHST approach to statistical testing, Chau's plan did not explicitly include specific provision for nor thinking about null models.

Given these results, one might ask whether these six students truly saw the generation of summary statistics or histograms as necessary steps in a plan for a statistical test, which is primarily predicated upon the comparison of a sample estimate to a null model. The answer to this question is unknown, as students were given a prompt in the form of a research question, and not instructed specifically to omit any steps they might normally take on the path to conducting a statistical test – they were left free to enact whatever plan they thought appropriate. However, as far as thinking about the statistical

nature of tests, i.e., the specification of a null model (or a bootstrap dot plot, depending on the approach), some students did not include a plan for the statistical aspect of a test at all. Of the six students, only Tal and Jaci explicitly incorporated some notion of probability and uncertainty in their plan for a statistical test, both in the form of the bootstrap dot plot.

5.1.2 The Nature of Graduate Students' Monitoring When Conducting Statistical Tests

A likely casualty to the natural progression of memory decay in the seven months since the six student participants completed EPSY 5261 and their participation in this study, the students' ability to monitor their process for completing a statistical test was poor.

For example, Aan's monitoring seemed to entirely consist of a comparison between 'what the means say' and 'what the *p*-value says' in terms of whether there was a difference between groups. Similarly, Kei's monitoring seemed to be based on Kei's visual memory, with Kei only being able to tell whether something 'looked wrong' but without any greater specificity.

Ade's monitoring was slightly more advanced and focused on the distribution within each group. In particular, Ade was sensitive to differences in shape, location as quantified by the mode, and spread as quantified by the range and sample standard deviation. Chau also focused on the distribution within each group, but the purpose of Chau's monitoring was for the verification of the normality assumption of the *t*-test. The rest of Chau's monitoring also focused on assumption verification, with Chau noting the study design, sampling strategy, and sample size, en route to conducting a *t*-test.

Tal's monitoring also focused on the sample distributions, but like Ade and unlike Chau, Tal's monitoring was to compare the distributions. However, unlike Ade, Tal

specifically contextualized the differences, noting differences in the ranges, the IQR, and even computing the percentage difference between the means of each group. Tal also spent time examining the null model (although Tal incorrectly thought it was a bootstrap dot plot), examining its shape, center, and spread. Tal paid particular attention to the upper and lower bounds of the middle 95% of the distribution, comparing these limits to the information obtained from comparing the means in each group.

Jaci, like Tal, also focused on the upper and lower bounds of the middle 95% of the null model (and similarly incorrectly thought it was a bootstrap dot plot). However, as opposed to Tal, Jaci focused on contextualizing each bound separately and noted whether they both had the same practical and contextual interpretation.

5.1.3 The Nature of Graduate Students' Evaluating When Conducting Statistical Tests

Even more so than monitoring, students' evaluating seemed to be a likely casualty of the natural progression of memory decay in the seven months since the six student participants completed EPSY 5261 and their participation in this study. For example, Kei admitted to almost entirely forgetting how to evaluate progress towards completing a statistical test. Similarly, Aan, due to confusion in remembering how to interpret *p*-values, changed their strategy to focus almost entirely on the estimate of the means of each group, which Aan felt comfortable contextualizing and interpreting. As opposed to Kei and Aan, Jaci was singly focused on generating a bootstrap dot plot to draw inference. Therefore, Jaci's evaluating was essentially non-existent, as it did not seem as if any of the other analyses Jaci ran led to any change in this central plan for conducting a statistical test, nor would they have regardless of the results.

Tal's evaluating, like Aan's and Jaci's, seemed to focus on the estimates of the means of each group, as well as the confidence interval and bootstrap dot plot. However, in comparing these estimates to the research question, Tal's strategy would only change if results were not "sensible". When this was the case, Tal would first check the data to ensure that there were no computational errors and then Tal would question the governing theory and the research question. Therefore, Tal's evaluating seemed to lead to changes in strategies in terms of thinking about the real world, rather than changes in strategies for specific statistical operations.

Compared to Kei and Aan, Ade's evaluating seemed to be a little bit more advanced and focused on the sample distributions. Specifically, Ade appeared poised to simply compare the means in each group to draw a conclusion, until Ade saw the differences in the sample standard deviations. This then caused Ade to generate histograms. Additional differences in the distributions between each group then led Ade to suggest that a statistical test ought to be performed.

Chau's evaluating entirely stemmed around the verification of the assumptions of a *t*-test. For example, during the *statistical testing task*, Chau determined that the VSE problem was not experimental in nature. This led Chau to evaluate their strategy in terms of utilizing a *t*-test. Eventually, Chau settled on using SBI methods, deciding that a *t*-test could not be used if the study was not experimental, but that a bootstrap dot plot could still be generated and utilized for making statistical inferences.

The fact that all six students evaluated their statistical tests differently begs the question as to whether there was consistent and specific instruction on how to monitor and evaluate statistical tests in EPSY 5261. Unfortunately, it may simply be that monitoring

and evaluating skills are acquired through experience, perhaps only with a level of experience unattainable within a single semester. Alternately, it may be that monitoring and evaluating are distinct skills that should be explicitly taught and practiced in the classroom and may have been underdeveloped in EPSY 5261.

Either way, these results seem to indicate a need for further research on the development of students' monitoring and evaluating when conducting statistical inference tasks as well as the identification of sensitive and specific research methods to distinctly and reliably measure differences in students' monitoring and evaluating proficiency, separate from their planning.

5.2 Research Question 2: Do graduate students think about null models when conducting statistical tests, and if so, how?

It appears that the six participants generally did not think about null models when conducting statistical tests. Even when presented with null models in simulation-based software tools, the participants erroneously conflated the null model with a bootstrap dot plot based on the observed sample. Furthermore, their thinking about these sampling distributions was inconsistent with the hypothetico-deductive approach to testing. Instead, the participants overwhelmingly thought about these sampling distributions in a manner consistent with the likelihood approach to statistics, with the bootstrap dot plot identifying which parameters were likely to be true given the observed sample statistic.

Specifically, two of the six participants, Aan and Chau, almost entirely focused on *p*-values, not being able to articulate what null models were, nor the role null hypotheses played in testing. Neither Aan nor Chau seemed to explicitly think about sampling variability, let alone null models. Furthermore, both Aan and Chau misconstrued the

randomization dot plot in StatKey for a bootstrap dot plot, essentially mischaracterizing a null model as something akin to a likelihood function.

Another two of the six participants, Tal and Jaci, did explicitly think about sampling variability, but not in terms of a null model. Both Tal and Jaci misconstrued the randomization dot plot in StatKey for a bootstrap dot plot, similar to Aan and Chau. However, both Tal and Jaci explicitly thought about the relative likelihood of parameters based on this distribution, even using the term "likelihood". Thus, while technically incorrectly using the randomization dot plot and technically thinking incorrectly from the standpoint of the classical school of statistics and its hypothetico-deductive approach to theory testing, Tal and Jaci both explicitly acknowledged and thought about sampling variability, albeit unwittingly in a manner akin to the likelihood approach to statistics.

The other two participants, Kei and Ade, were able to connect the null hypothesis to the center of the randomization dot plot in StatKey. However, the extent to which Kei and Ade understood the randomization dot plot as a null model, or the extent to which they thought about this null model across all testing tasks, is somewhat unclear. For example, Kei seemed to suggest that one need only to "think about it [the estimated difference]" to answer a research question, omitting the need for a null model in the logic of statistical testing. However, the fact remains that Kei and Ade were able to connect the null hypothesis to the null model, through its center.

These results, that most participants did not think about null models and the two that did only barely did so and with some difficulty, are not surprising. Previous research has found that students and in-service teachers overwhelmingly think about statistical tests not in terms of the process of the test (of which null models are the key), but rather in terms

of the product of the test, i.e., a *p*-value (e.g., Justice et al., 2018; Noll et al., 2018b) – this is akin to Aan's and Chau's thinking.

Especially when using software such as R, a product-based thinking about statistical tests does not require one to think about a null model at all. One needs only to identify the correct test and utilize the correct function to obtain the all-important *p*-value.

However, this is not true of simulation-based software such as StatKey. There, one physically generates a null model. Yet, four of the six students in this study did not recognize that the null model is based on a null hypothesis. Conceptually, randomization resampling based on a null hypothesis and bootstrap resampling based on the observed sample statistic are largely the same, taxonomically speaking from the perspective of different types of simulations and simulators. Compounding the issue of their conceptual similarity is that in StatKey, the user interface is nearly identical between these two use cases. It is incumbent upon students to remember to think about the center of the sampling distribution to determine whether they are working with a bootstrap distribution or a null model. In this study, the participants were mostly unable to recall this important distinction.

5.3 Limitations on Inferences and Conclusions

The purpose of this study was to describe graduate students' thinking about statistical tests and null models. Therefore, one must consider the study tasks' ability to elicit students' thinking, the ability of the combination of data sources and artifacts to capture a trace of students' thinking, the ability of the analytical approach to detect relevant aspects of students' thinking, and finally the limitations of inferences that can be made based on the individual students who participated in this study and the manner in which they were recruited.

5.3.1 Limitations Based on the Tasks Utilized

This study utilized multiple tasks to elicit participants' thinking. This was intentionally done as the current body of literature on students' statistical thinking is mainly based on single modality studies, typically in the form of either task-based interviews (e.g., Justice et al., 2018), group observations in a classroom (e.g., Noll et al., 2018b), or the analysis of students' written work (e.g., Frischemeier & Biehler, 2013). More importantly, there are only a few surveys designed to distinctly measure students' statistical thinking, as opposed to their reasoning, conceptual understanding, literacy, or any other construct, and it is an open question as to whether these assessments are able to distinctly measure statistical thinking. Nevertheless, thinking and reasoning are well defined distinct cognitive processes.

However, not all of the procedures and materials as designed in this study were equally useful in obtaining traces of students' thinking. For example, the *statistical testing task* presented open-ended research questions that, while requiring the use of a statistical test, did not mandate it. Indeed, nearly all participants computed summary statistics, which, while certainly a useful part of conducting a statistical test in general, is not core to the logic of statistical testing per the philosophy of statistics. Were these computations core to students' logic of statistical testing, or were they simply a part of the procedures students were taught that are now procedurally associated with the logic of statistical tests? The answer to this question is unclear from the data collected in this study. Thus, while the open-ended nature of these tasks was designed to allow for an interpretivist analysis of students' thinking, the general nature of these tasks was perhaps insufficiently specific to deeply probe and elicit students' thinking about the core aspects of statistical testing.

Nevertheless, relative to prior research studies, these tasks did place additional focus on testing, and specifically thinking about the null model. It may simply be that students do not remember how to think about null models without specific and recent training.

Additionally, some of the statistical test results presented to participants in the *statistical testing interview* contained intentional errors. This was done to elicit aspects of students' monitoring and evaluating. In identifying an error, participants would necessarily have had to articulate their monitoring and evaluating, as they identified what information they checked and how they determined something was amiss. Even a pause to consider something that may seem amiss would have been an insightful trace of students' monitoring, akin to preferential looking studies in which extended gaze is typically the basis of an inference that participants were surprised to see a particular stimulus.

However, none of the participants in this study identified any of the intentional errors in the stimuli in the *statistical testing interview*. Even when participants were queued by the researcher to examine the piece of erroneous information, they failed to identify that the output was impossible. One participant, Tal, even commented that the researcher might place intentional errors in the stimuli, but still failed to recognize the errors. Therefore, for these participants and the level of statistical thinking they were able to achieve in this study, identification of errors is perhaps too challenging a task. This is akin to writing test items that have too high a difficulty level for participants, leading to a floor effect in participants' scores, which then occludes the analysis of the construct of interest. Furthermore, participants were not instructed to look for errors. Therefore, it may be inappropriate to expect participants to exhibit the conjectured cognitive processes that the intentional errors were meant to elicit. Further research efforts designed to utilize intentional errors in the

study of statistical thinking should thus include explicit instruction to participants to look for such errors in separate assessment items.

It is worth noting that one of the pilot participants, an experienced statistics educator, was able to correctly identify the intentional errors without any specific prompting or priming by the researcher. Thus, while such tasks may be appropriate for experts, they are likely too difficult for novices. Yet, stimuli with intentional errors may still be useful in instruction or in the assessment of intermediate to advanced students.

Because this study's participants were unable to identify that they were looking at errors, they may have inadvertently abstracted incorrect information from the stimuli as they attempted to recall statistical knowledge. Therefore, results from the *statistical testing interview* should be interpreted with this fact in mind, and that any trace of a participant "incorrectly" thinking, especially with regards to *p*-values, in that task and in the *concept mapping task* should not serve as the basis of any inferences about students' thinking.

Finally, it must be noted that each participant took approximately 90 minutes to complete all tasks as part of this study. It is possible that the participants felt some level of cognitive fatigue by the end of the study. Therefore, this possible fatigue must be kept in mind, especially when interpreting results from the *statistical testing interview* and the *video-cued interview*, which both came after the *concept mapping task* and the *statistical testing task*.

5.3.2 Limitations Based on the Data Collected

The multiple modalities of data collected from each of the tasks in this study were designed to aggregately serve as a robust method of capturing traces of participants' thinking, particularly in terms of their monitoring and evaluating. This was intentionally

done as the debate on the extent to which think-aloud procedures interfere with thinking is one without consensus. Indeed, some participants seemed able to flawlessly elucidate their thinking, such as Tal, while others required the researcher to prompt them to fully elicit their thinking, such as Aan. With this in mind, it is important to note that there were many moments in which the researcher could have, and perhaps should have, asked the participants to explain their thinking out loud, rather than sit in silence. This, perhaps, would have been a more faithful application of the think-aloud procedure. However, because there was the gaze recording, which would serve as a trace of thinking even in moments of silence, as well as the *video-cued interview*, in which the researcher could retrospectively ask the participant to explain their thinking, the researcher chose to make the decision to prompt the participant to think aloud conservatively. Therefore, it is possible that some key moments of participants' in-the-moment monitoring were not captured by the audio and video recordings. With a trace of these moments of monitoring only captured by the gaze recording, these moments were not able to be reliably triangulated. Thus, there may be gaps in the description of each participant's thinking, especially in terms of their monitoring. Thus, absence of evidence of specific types of thinking should not be construed as evidence of an absence in participants' thinking during this study's tasks.

Additionally, and contributing to the potential misalignment of the difficulty of these tasks with participants' abilities is the fact that, participants were explicitly instructed not to prepare or review any materials prior to participating in this study. The purpose of this instruction was to provide a way to identify the most memorable aspects of statistical thinking, which might then serve as an anchor in the future design of activities and lessons. However, it was clear that the participants took some time to settle and slowly recollect

how to do statistical analyses. Combined with the social pressure of wanting to 'do well' while sitting next to the researcher, this instruction to not prepare may have negatively affected the measurement and inferences of participants' thinking. Furthermore, it was not ecologically valid, as these participants would have utilized materials and notes had they been assigned a statistical testing task outside of the confines of a research study. Therefore, results from all tasks should be interpreted with this fact in mind, that participants had completed EPSY 5261 seven months prior to participating in this study and that they did not review any study materials before beginning the study tasks.

Finally, the largest limitation is one of data collection errors. For both Jaci and Chau, the audio recording of the *statistical testing interview* was corrupted and thus unusable. However, in both cases, the screen recordings and the researcher's observer notes were used to support analyses of Jaci's and Chau's thinking. Similarly, for Ade, the video and audio recording of the *video-cued interview* failed. However, the researcher recognized this immediately after the conclusion of the *video-cued interview*, and thus the researcher, the research associate, and Ade all immediately wrote down a summary of the discussion that had occurred during the *video-cued interview*. These notes, in addition to the researcher's observer notes, were used to support analyses of Ade's thinking.

5.3.3 Limitations Based on Researcher Reflexivity

Beyond limitations of study design and data collection, it is imperative in qualitative research to consider the inherent biases of the researcher when conducting analysis and the analytical approach taken.

First and foremost, the researcher was one of the instructors of EPSY 5261 in the Fall 2021 semester from which three of the six participating students were recruited. Not

only does this create the potential for bias in the analysis, but it also creates a bias on the part of the participants, who were asked to sit next to their instructor in a research setting while completing the study tasks. This social desirability bias was also likely present for the other three participants as well, as their instructor, while not the person sitting next to them during the data collection, was also involved in this study.

With the potential for bias in the analysis, as participants' performance could potentially be construed as a reflection of the quality of the researchers' teaching, what steps were taken to mitigate this bias? First and foremost, it is important to state the potential desirability of results on the part of the researcher. Prior to data collection, the researcher assumed that students would not remember much in terms of what they had learned in EPSY 5261, but rather than take a deficit model approach to the analysis, the researcher's goal was to describe what the students did remember. As there have been no previous studies examining graduate students' thinking about statistical tests with a distal measure similar to the seven-month delay between the completion of EPSY 5261 and students' participation in this study, there was no specific desirability in terms of what students should remember, beyond the learning goals of the course. The primary objective of the research was thus first and foremost to simply describe what the participants did, to formally enter into the empirical record evidence of students' thinking several months after completing an introductory level course utilizing an SBI curriculum.

However, what the teaching team of EPSY 5261 would have hoped students remember was that they would remember what a null model was, where it comes from, and what it represents, as discussed in a meeting in 2019 (well before conceptualization of this study) and shown in Figure 23. While the researcher of this study was a part of that

discussion, the assumption prior to conducting this study was that participants likely would not remember a null model, as it was not an explicit focus of instruction in EPSY 5261. Thus, the concept of a null model could have easily been lost among the other topics covered in the course, and in the time that passed since the completion of the course and this study's data collection.

To ensure the credibility of the results of the analysis, the study design employed several strategies, as detailed in Chapter 3. However, one potential weakness of the design is that there was not a second researcher who examined the data, and thus no comparison of the reliability of the codes that the researcher extracted from the data. Nevertheless, some of the codes and relevant moments were discussed directly with participants as part of the *video-cued interview*, adding a measure of credibility to the findings of this study.

5.3.4 Limitations on the Generalizability of Results

With an intentionally recruited sample of six students, few if any results can be generalized from this study. Indeed, generalization of results was not a primary objective of this study. However, it is also important to think about the characteristics of these students, the courses they took, and the institution at which they took the courses.

All students completed EPSY 5261 in the fall of 2021, an introductory level statistics course at the master's level, at the University of Minnesota Twin-Cities. There were two instructors of in-person sections that semester. Both were experienced statistics educators, having taught the course many times previously, and being familiar with historical and current trends in statistics education research. The course utilized the *Statistics: UnLOCKing the Power of Data* curriculum (Lock5; Lock et al., 2021),

accompanied by activities designed by the teaching team at the University of Minnesota over the previous decade.

However, there were some variations in how each instructor taught the course. For example, one instructor presented study design diagrams as a framework for thinking about study design and the design of simulators. The purpose of these study design diagrams was to focus on data generating processes as well as to address potential misconceptions related to simulation as identified by Brown (2021). Of the three students who participated in this study who were taught these study design diagrams, only one mentioned these diagrams in the study. Kei utilized study design diagrams when drawing a concept map for the logic of a statistical test. To what extent did instruction with these diagrams and the fact that Kei remembered these diagrams affect Kei's thinking when conducting statistical tests? The answer is likely not all that much. Kei's thinking while conducting statistical tests did not seem to follow the specific plan or framework of the study design diagrams. Instead, Kei's thinking seemed to focus on the software Kei was utilizing and procedural memory. Kei's thinking thus was different even from Kei's two classmates who also participated in this study, and was perhaps most similar to Ade's, who was not in the same EPSY 5261 section as Kei. What makes Kei's and Ade's thinking similar? It seems to be the relative recollective ability demonstrated in the tasks that make Kei's and Ade's thinking most similar. Both expressed doubt in completing the tasks. Both employed specific plans for conducting a statistical test but struggled in monitoring and evaluating their plan. Both were able to eventually remember that the center of the randomization dot plot is related to the null hypothesis. Perhaps most importantly, both had very little prior statistics

experience before completing EPSY 5261 and very little experience with statistics since having completed EPSY 5261.

This last point is important. For all the individual differences that may manifest in the classroom, and all of the environmental differences, such as different instructors or different curricula, what seems the simplest explanation for the similarity in Kei's and Ade's thinking, along with all the other participants, is the simple fact that we humans ubiquitously forget information and this forgetting is only abetted with reinforcement. The fact that Kei's and Ade's thinking is similar is a possible testament to the efficacy of the Lock5 curriculum and the consistency between the instructors of EPSY 5261 that both Kei and Ade had. Perhaps, more generally, it is a testament to the SBI approach to teaching statistical inference, which after all aimed to place the logic of statistical testing at its core. Therefore, while there are certainly many factors that affect students' learning and students' performance, it may simply be that time vanquishes all, and in longitudinal studies of students' statistical thinking, individual and environmental effects are somewhat muted relative to the effect of individual differences and environmental effects on proximal studies and measures most common in the statistics education literature.

5.4 Implications for Teaching

Given the findings of this study, there are two main questions with regards to the implications for teaching that arise: (1) Should one-and-done students take an SBI course? and (2) Should one-and-done students take a course based on the classical school of statistics?

5.4.1 Should One-and-Done Students Take an SBI course?

SBI courses aim to put the logic of statistical inference at their core. With regards to statistical testing, it is the null model that occupies the key role in the logic of a statistical test. SBI software applications all display null models and it is with these null models that students directly interact. However, in this study, most participants conflated this null model with a bootstrap dot plot at one point or another. Does this mean that SBI curricula have failed in their task?

Each SBI curriculum is different and they also differ in the degree to which students interact with null models. For example, in the ISI and Lock5 curricula, students generate null models only by clicking a 'simulate' button, but the rest of the characteristics of the model is specified by simply knowing which link to click (e.g., test for a difference in means, test for a single proportion, etc.). In the CATALST and CourseKata (Son et al., 2021) curricula, students build software models that generate null models. In CATALST, students build TinkerPlots models, and it is the design of these models, meant to replicate the data generating process of the real world, that leads to the specification of the shape, center, and spread of the null model. Similarly, in the CourseKata curriculum, students build models in R, before resampling data to produce sampling distributions. It is currently unclear to what extent these different curricula and the different nature of students' interaction with models (and the null model in particular) affect students' thinking, both within the course and well after the course ends.

Furthermore, the six participants in this study all seemed to remember how to think about bootstrap resampling quite well – it was null models that proved difficult. Therefore, it stands to reason that, at least with regard to the logic of estimation, this curriculum was effective in developing these participants' statistical thinking. It is more likely that the

participants' difficulties in thinking about null models in this study are not due to any failure of the SBI curriculum, but rather due to the inherent complexity and difficulty in thinking about null models, null hypothesis significance testing, and more generally the hypothetico-deductive approach to confirmation theory.

However, whether this course was effective in its goals is only one consideration that should be made given that these students were mainly one-and-done statistics students. With only one semester of statistical training to work with, what should instructors and program directors focus on teaching for such graduate students? Are the benefits of teaching an SBI course, focused on developing a conceptual understanding of statistical inference and its logic, worth the opportunity cost? The answer to this question may indeed be 'no'.

This point is perhaps most clearly elucidated by Kei, who commented that, "For the class, I think having actual papers, and having more time to apply it to things we're actually reading and doing would have been beneficial. Like, I get the point of simulation, but I'm like … especially for intro to stats people, we're not going to be doing any of that stuff for a minute, and so I think it's more important that we know how to read an article and know what to take from an article than to do weeks on simulation. Cuz we're not going to be working with our own data for a minute. So, I think it's more important especially for 1st and 2nd years, and even master's students, to know how to read an article and know what to look for and how to critique it" (Appendix P03-E, 05:07 – 06:00).

This point, that the opportunity cost for one-and-done graduate students is too high when spending their one semester on introductory SBI methods, is made even more poignant when combined with the fact that recent research has shown that prior experience

is a larger determinant than curricular differences in students' conceptual understanding in introductory statistics courses (Chance et al., 2022). This implies that there is, at least in the short-term, little curricular effect on students' conceptual understanding, and thus that the choice of curriculum might be made on other considerations, such as practical training in reading a literature, as suggested by Kei.

Additionally, posterior experience, that is, whatever experiences occur after students complete an introductory level course, also appears to play a large role in students' statistical thinking skills. Of the six participants in this study, Tal and Chau were the two participants with the most experience in conducting quantitative analyses both before and after having completed EPSY 5261. While Tal and Chau made similar errors in thinking about null models as the other four participants, both were particularly comfortable in thinking about data, in computing and interpreting summary statistics and data visualizations, and in thinking about confidence intervals. For one-and-done students, the hard-earned conceptual understandings obtained in SBI curricula, without reinforcement, seem likely to quickly fade. Even for Tal and Chau, perhaps because the way statistics is practiced is typically not simulation based, both reverted to thinking about 'means' and '$p < .05$' in R. This might be due to the nature of statistical practice in their fields, what they are most exposed to, and the practices that are most reinforced. As the aphorism goes, 'Out of sight, out of mind'.

Therefore, despite the evidence that SBI curricula seem to do at least as well as consensus curricula in developing students' conceptual understanding of statistics in end-of-term assessments, for one-and-done graduate students, it may be that the opportunity cost of using an SBI curriculum, at least among the common SBI curricula that currently

exist, may be too high a cost. Instead, these students' short time in a statistics classroom may be better spent focusing on more practical and translational considerations related to statistical thinking.

5.4.2 Should One-and-Done Students Take a Course Based on the Classical School of Statistics?

It is well known that the logic of statistical testing within the classical school of statistics is confusing to many students and often appears erroneous even in textbooks (Nickerson, 2000). The students participating in this study were no exception. Most struggled to think about null hypotheses and *p*-values, and nearly all struggled to think about the randomization dot plot (i.e., the simulation-based null model). However, nearly all students correctly interpreted and thought about confidence intervals. Furthermore, most students thought about the bootstrap distribution in a manner akin to a likelihood function quite correctly and intuitively, albeit unwittingly. Even though statistics educators have spent decades of work attempting to develop new methods to clarify thinking in the hypothetico-deductive approach required of the classical school, students still appear to struggle with the concepts. Given evidence of seemingly effortless and fluent thinking by students consistent with the likelihood-based approach, why not adopt a different school of statistics?

Indeed, the Bayesian school of statistics has, since the start of the 21st century, been gaining widespread popularity. This is especially true in computationally complex fields such as genomics which have greatly benefited from Bayesian computational approaches. Philosophically speaking, the Bayesian approach to confirmation is also a more widely accepted approach when compared to the hypothetico-deductive approach to confirmation.

These two facts, when combined with the seemingly intuitive manner in which students in this study thought in terms of likelihood, provide a compelling argument to focus on the likelihood function as the basis for statistical inference in introductory courses for one-and-done students and not null models. As Bayesian posterior probabilities are proportional to the product of prior probabilities and the likelihood function based on the observed data, a likelihood-based framework might be easily scaled up to incorporate the notion of prior probability.

However, computing posterior probabilities are computationally intense, relative to the introductory SBI curriculum. Therefore, it seems appropriate to instead only focus on the likelihood function, as generated via bootstrap resampling based on some observed sample, and to use this likelihood function to extract a credibility interval, and in general, to reason about the relative likelihood of various parameters, given the observed evidence. Tests could then simply take the form of the likelihood ratio test, with a focus on Bayes factor, rather than *p*-values.

It is important to note that while such an approach might serve to develop students' thinking about uncertainty and probability in statistical tests, such a course based on the likelihood function would be misaligned with historical and current practice, which predominantly favors the classical approach to statistical testing and NHST. In essence, fluency and expertise in thinking about statistical tests through a likelihood function may not serve the needs of one-and-done students. Therefore, it is important to consider whether, compared to current SBI curricula, a course grounding statistical testing with the likelihood function could serve as a foundation to subsequently develop students' thinking about null models in a hypothetico-deductive approach to statistical testing. Nevertheless,

as the 21st century progresses, Bayesian methods appear to be promulgating across science, and thus, preparing students to think about likelihood functions in addition to hypothetico-deductive approach may serve their future needs in addition to their current needs.

Given the evidence observed from the six students who participated in this study, successfully developing students' statistical thinking with such an introductory likelihood-focused simulation-based curriculum appears attainable. For example, Tal specifically commented upon the varying degrees of likelihood with which a parameter may be true.

Furthermore, and in a departure from the study tasks and stated research questions, after Jaci completed the *Concept Mapping Task*, the researcher very briefly explained prior and posterior distributions – prior distributions as 'what you think walking in before you have collected any data' and posterior distributions as 'what you think after collecting data'. Then, the researcher presented a drawing of a prior and posterior distribution to Jaci, using the context of home prices in New York from *the Statistical Testing Interview* (see Figure 24). Jaci was then asked to interpret the distributions relative to the question "Is the average home price equal to $300,000 in New York?". Jaci intuitively thought about the posterior distribution, commenting that "The posterior is obviously higher than the supposed hypothesis" (Appendix P04-A, 05:18 – 05:22).

Jaci almost assuredly did not fully appreciate the intricacies of posterior distributions, given that Jaci had never received instruction on this topic. However, the fact that Jaci was able to think about the posterior distribution and draw a conclusion in an accurate manner without instruction is an indication of how Jaci thinks about bootstrap dot plots and sampling distributions in general. Indeed, many researchers have argued for the relative intuitiveness of Bayesian statistics. Importantly, Jaci's thinking provides some

empirical evidence that students may indeed find thinking about posterior distributions easier than they find thinking about null models, corroborating researchers' claims. Given the philosophical and practical advantages in addition to this potential pedagogical advantage, it seems a worthy endeavor to explore simulation-based Bayesian introductory statistics courses as a means for teaching statistical inference to one-and-done graduate students.

5.5 Implications for Practice

There are two specific recommendations for practice based on the performance of the six students participating in this study that are particularly noteworthy. First is that practicing statisticians should explicitly provide null models for all statistical tests they conduct. Second, practicing statisticians and software developers must carefully consider the user interface of their software, as students' statistical thinking may be affected by the specific manner in which information is presented.

5.5.1 The Explicit Specification of Null Models

As the proverbial aphorism goes, 'out of sight, out of mind'. For the six participants in this study, the fact that none has seen a null model since they completed EPSY 5261 is perhaps the simplest and most influential factor that can explain their difficulty in thinking about null models, and in general, their thinking about null hypotheses and *p*-values. Indeed, the few students who had done some quantitative analyses or read papers with quantitative results since completing EPSY 5261 seemed to be quite fluent in interpreting means and typing it to the real-world context, as well as in interpreting confidence intervals. These students even were fluent with the confidence interval approach to hypothesis testing, in which one only considers whether a particular value for a parameter

is contained within the confidence interval to judge whether it is a plausible value for the parameter of interest, 'rejecting' the value as implausible if it is not so contained. Despite this method being equivalent to '$p < .05$' thinking, the participants in this study did not see these two methods as equivalent and struggled when thinking about *p*-values. Why the participants were generally more comfortable with confidence intervals than *p*-values is an open question. At the very least, regardless of the manner in which students are taught about each (whether it be through simulated distributions or by examining the *t*-distribution), no sampling distribution is typically included in statistical software output. As probability distributions are the engine of statistics, it seems to be that one semester of introductory statistics is not enough to clearly entrench these distributions in students' minds, such that they coherently think about them when faced with statistical inference tasks. This is not surprising, as sampling distributions are famously one of the most difficult concepts in statistics to understand. Yet, they are the core of statistical inference. Thus, within the span of one short semester for one-and-done students, instructors must ensure students abstract and generalize the concept of a sampling distribution with such fluency that they are able to routinely apply it. This is no easy task, especially given that much of a typical semester is spent on other topics and not solely on statistical inference and sampling distributions. It is likely that such students require additional scaffolds as they enter statistical practice.

Thus, when considering statistical practice, it bears asking "Why only write 'H_0: $\mu 1 - \mu 2 = 0$'"? This does not meet the criteria of fully specifying a null model. Implicit for experts is that the parameter "$\mu 1 - \mu 2$" will be estimated by the sample statistic of the difference in sample means between two groups, that the sample statistic's distribution will

be normally distributed per the Central Limit Theorem, and that the spread of the distribution will be estimated based on the underlying variation observed within each group in the sample and the known sample size, as per the Central Limit Theorem. Do students implicitly understand these two additional required criteria when they see "H_0: $\mu 1 - \mu 2 = 0$"? The answer, based on the researcher's own experiences and beliefs, is likely no. Therefore, one potential practice that might develop the habit of graphically specifying null models for all statistical tests they conduct, which would additionally serve as a scaffold for students in thinking about statistical tests, reinforcing their thinking about null models beyond their time in statistics classrooms.

However, it should also be noted that curricular design can also support the development of the habit of mind in thinking about null models, which could, in conjunction with practice, serve to develop and reinforce students' thinking about null models. For example, one-and-done courses could be designed solely around developing students' understanding of and thinking about statistical models used in real-world studies. In such a course, always asking students to identify the null hypothesis statement for each statistical model or test as well as the two additional required criteria to fully specify a null model might also serve to develop students' thinking in a manner consistent with the core logic of statistical testing. Such a course, utilizing real-world studies, would also support the reinforcement of students' thinking about statistical testing once they enter statistical practice, so long as statistical practice similarly emphasized the explicit and complete specification of the null model in statistical testing.

5.5.2 The Careful Consideration of Statistical Software User Interfaces

The aphorism ‘out of sight, out of mind’ is a useful reminder for practicing statisticians to consider what pieces of information should be explicated in statistical communication, but its inverse, ‘what is in sight is in mind’ must also be considered. Specifically, the difference in layout of the various statistical software that is utilized seemed to play a role in this study’s participants’ thinking.

For example, Tal, who has experience with SPSS, was seemingly annoyed by the way in which R presented information. Tal then commenced to utilize an SPSS-based schema to think about the output R was providing. Similarly, Kei was particularly sensitive to noticing pieces of information that were intentionally deprecated in StatKey output in the *statistical testing interview*.

Students’ schemas for statistical thinking being dependent on the software layout is a prediction of a strong theory of instrumental genesis. In Kei’s example, it is unclear whether Kei saw R and StatKey as equivalent statistical tools. Certainly, Kei’s thinking was quite different when using each of the two software tools, even beyond the procedural differences that would be expected from differences in the design of each software tool.

If students’ thinking is dependent on, or even moderately correlated to, which software they utilize, then this dependence may inhibit transfer of students’ statistical thinking to other problems and certainly other software applications. However, it provides an interesting piece of information for instructional designers and for practicing statisticians. If novices’ statistical thinking is intrinsically tied to the manner in which software tools present statistical output, then not all software tools are equivalent. Practicing statisticians should privilege those software applications that present information in a manner most consistent with the core logic of statistical inference.

5.6 Implications for Research

As this study utilized a unique combination of methods and sources of data, it bears considering the relative value of each of these sources and tasks, as might inform future research on students' statistical thinking. However, first and foremost, this study shows the need for additional longitudinal studies of students' statistical thinking, or studies with distal measures of students' statistical thinking.

5.6.1 The Value of Longitudinal and Distal Studies of Students' Statistical Thinking

This study was one of the first to study students' statistical thinking several months after they have completed an introductory level course. While some students, particularly statistics majors at the undergraduate level, are likely to take additional statistics classes that will build on material in their introductory level course, it still bears considering what these students remember well after the completion of an introductory level course. It is well known that individuals forget information over time. The learning objectives for a statistics course, or any course, are rarely designed with the intention that students need not remember the information after the course has been completed. Therefore, what students remember and how they remember it is an important piece of information that can inform statistics education research, especially at the post-secondary level. For example, the fact that the participants in this study were able to think about bootstrap dot plots and confidence intervals confidently and correctly is important information for instructors in determining how long to spend on various topics, as well as for researchers studying course sequences and the efficacy of various scaffolds to support student learning. Furthermore, this information, when combined with previous studies examining students' thinking about estimation and their thinking about testing, can help to support inferences and theories

about students' statistics education. Therefore, the first implication for research from this study is that statistics education researchers should conduct more longitudinal studies, or studies with distal measures of students' statistical thinking.

5.6.2 The Relative Value of the Tasks Used in this Study

This study utilized four different tasks to elicit participants' thinking. Thus, it bears considering which tasks helped to support inferences about each of the different aspects of thinking: planning, monitoring, evaluating.

For making inferences about participants' planning, the *concept mapping task* and the *statistical testing task* were particularly helpful. The *concept mapping task* provided participants an opportunity to say what they would do and then the researcher could observe what they actually did in the *statistical testing task*. Together these tasks helped triangulate participants' planning. Having both tasks was particularly helpful, as there was some small misalignment between the planning across tasks for most of the participants. This may have been due to specific intricacies of each statistical software tool or may also be due to participants' evaluating and revisions to their planning that occurred in the *statistical testing task*.

For making inferences about participants' monitoring, the gaze recordings from the *statistical testing task*, the *statistical testing interview*, and the *video-cued interview* were particularly helpful. Not all participants clearly articulated their stream of consciousness by thinking aloud in the *statistical testing task* and thus the gaze recordings proved an excellent resource in terms of capturing students' in-the-moment thinking. Furthermore, the use of the gaze-recordings in the *video-cued interview* was particularly helpful in retrospectively prompting participants to reflect upon and elucidate their thinking during

the *statistical testing task*. Aside from simply seeing a screen recording of what they had done in the *statistical testing task* or hearing what they had said from an audio recording, seeing their gaze recording provided participants an opportunity to explain even their most confusing moments in the *statistical testing task*, which most often aligned with students' monitoring. Especially given the seven-month delay between the completion of EPSY 5261 and their participation in this study, most participants spent a substantial amount of time 'scanning' the screen, which was readily observed by examining the gaze recordings. Thus, using the gaze recordings in the *video-cued interview* provided an opportunity for participants to retrospectively comment on all aspects of their thinking, including these moments of 'scanning' and monitoring. Additionally, the *statistical testing interview* provided an interesting insight into how the participants processed the information provided by statistical software tools reporting the results of statistical tests. This information processing most closely aligns with the monitoring subcomponent of thinking. While the difficulty level of the stimuli used in this study was perhaps too difficult for these participants, this task, or similar tasks, again in conjunction with gaze recordings, may prove a useful tool in future studies of students' statistical thinking.

For making inferences about participants' evaluating, the *statistical testing task* and especially the *video-cued interview* were particularly helpful. In the *statistical testing task*, moments in which participants were evaluating were primarily identified by participants' articulation of changes to their original plan. With prompting by the researcher, participants were then able to articulate their thinking with regards to their evaluation of information. However, the *video-cued interview* provided an equally helpful measure of evaluating, if not more so. Perhaps because the six participants in this study were all graduate students

in the educational and psychological sciences, their ability to comment on and analyze their own thinking was uniquely enabled in the *video-cued interview*. In that task, the participants were able to take a step back from their in-the-moment thinking in the *statistical testing task* and provide insights that served not only as credibility checks, but also provided context and detailed explanations of their own thinking. Therefore, future research in students' thinking might benefit from the use of video-cued interviews, even if those interviews are conducted without gaze recordings.

5.6.3 The Relative Value of the Data Sources Used in This Study

This study utilized three primary sources of data across all tasks: audio recording, screen recording, and eye gaze recording. The use of audio and screen recordings is quite common in statistics education research, not only for the study of individual students' thinking but also for groups of students. However, the use of eye gaze recording is relatively novel in the study of students' statistical thinking. Thus, it bears considering whether the costs of collecting this source of data are eclipsed by its benefits.

The costs of capturing gaze recordings are considerable. The physical equipment alone can cost several thousands of dollars. Furthermore, utilizing the complicated equipment and its associated software also requires substantial training. This is particularly true for the calibration process, in which the headset that participants wear to capture their eye gaze must be physically adjusted. This study was only able to utilize these tools by the support and collaboration of a research associate who had access to eye-tracking equipment, was well-versed in the utilization of this equipment, and who was present to ensure that gaze recordings were successfully captured for all participants.

In addition to these costs to the researcher, we must also consider the costs to the study participants. While participants were financially compensated for their time, especially in part due to the invasive and perhaps uncomfortable nature of wearing the eye-tracking headgear, this may have still been uncomfortable for students. Although the participants in this study did not explicitly comment on any discomfort with regards to wearing the eye-tracking headgear, wearing the equipment may have affected their thinking. Perhaps because all six participants in this study are graduate students in the educational and psychological sciences, they seemed more interested and curious in their gaze data, rather than uncomfortable. Yet, this might not be the case for other participants.

Given these costs, what then were the benefits of the gaze recording data? The gaze recordings were particularly helpful in making inferences about students' monitoring, especially in cases where the students were not comfortable articulating a stream of consciousness. For example, Tal was particularly comfortable with the think-aloud procedure, explaining in detail their thinking. Thus, for Tal, the gaze recordings only seemed to serve as an additional source of data in the triangulation of Tal's thinking. However, this benefit should not be discounted. While Tal's planning could be triangulated through multiple tasks and sources, Tal's monitoring, and in general all the participants' monitoring, was much harder to capture. Especially for studies focusing on students' monitoring and evaluating, these gaze recordings might be an invaluable source of data to help triangulate their thinking, beyond what can be obtained via think-aloud procedures.

For participants who did not explain their thinking aloud as seamlessly as Tal, such as Aan, the gaze recordings were of a greater benefit. Aan stayed silent many times as they were processing information and thinking, and despite being prompted to think out loud,

still often stayed silent. While researcher queueing was able to elicit aspects of Aan's monitoring, these were not as faithfully in-the-moment as with Tal's explanations. Therefore, the gaze recordings for Aan served as the primary data source in terms of what Aan was monitoring, in what order, and how Aan eventually arrived at the conclusions that Aan did explain out loud. Thus, gaze recordings can be a vital source of data that is cognitively non-invasive, compared to think-aloud procedures, especially for participants who are not adept at explicating their stream of consciousness.

One limitation of the analysis of gaze recordings is that a researcher must be careful in thinking about the discriminability of the gaze recordings to identify particular aspects of thinking. That is, one must think not only about what aspects of participants' thinking might be inferred from a particular gaze pattern, but also what all possible gaze patterns might be produced by that particular aspect of participants' thinking. This analysis, in essence producing a two-by-two table of aspects of thinking and evidence in the gaze recordings, is consistent with modern approaches to video-based data and would be an essential step for a more thorough and in-depth analysis of the gaze-recordings as the primary data source (DeLiema et al., 2023). These analyses could be done for a single study participant, or ideally, across participants, to generate a more generalizable framework for the study of students' monitoring.

5.7 Conclusion

Might SBI curricula be able to support graduate students' development of an understanding of the core logic of statistical testing? How do graduate students who have completed a master's level introductory statistics course utilizing an SBI curriculum think when they conduct statistical tests?

To answer this question, Chapter Two of this dissertation first examined the history and philosophy of statistics, identifying that the key to the logic of a statistical test is the exact specification of the probabilities with which possible sample statistics may occur given some hypothesis to be tested (i.e., a null model). Next, the chapter examined current understanding of students' thinking within and after having completed an SBI introductory level statistics course, finding that SBI curricula appear to support the development of students' conceptual understanding of statistics, but there is a critical gap in students' understanding of null models.

In Chapter Three, the study design for this study was presented, describing the multi-modal method of capturing traces of students' thinking based on four tasks (i.e., the *concept mapping task*, the *statistical testing task*, the *statistical testing interview*, and the *video-cued interview*), as well as several data sources (i.e., audio recordings, screen recordings, eye gaze recordings, video recordings, and observer notes).

Chapter Four presented the results of the study, highlighting relevant traces of participants' thinking as they related to statistical testing and specifically thinking about null models. Results suggested that participants' planning was generally quite good, as most participants remembered what needed to be done to conduct a statistical test, although their plans were more general than those aspects specific to the statistical test itself (i.e., computing means, generating histograms). However, participants' monitoring and evaluating were generally quite poor, as they struggled to remember how to enact their plan and when they might need to revise or change their initial plan. Finally, results suggested that very few participants explicitly thought about sampling variability, let alone null

models, instead thinking about statistical tests only through their product, either in the form of the confidence interval or a *p*-value.

Finally, Chapter Five situated the findings of this study against prior literature. It also presented conjectures for possible ways to build off of this study's results, including directions for future research in terms of the utilization of various tasks and data sources to study students' thinking, directions for teaching in terms of the appropriateness of SBI curricula for one-and-done graduate students as well as the potential to focus more explicitly on null models and/or likelihood functions and, finally, directions for the practice of statistics, especially in terms of how null models are communicated and the role that software plays in reinforcing students' statistical thinking.

The cycle of science begins with observations of intriguing and perplexing phenomena that spurs theory generation and testing to explain the process behind those observations. These scientific theories are not judged by their truth, but rather by their ability to empower. How then does this study empower us? Learning is a complex process and learning statistics is no different. Historical difficulties in learning statistics have led to simulation-based curricula, which aim to promote students' learning of statistical inference. Current understanding of students' learning with these curricula was previously limited to studies conducted while students were still completing, or had just finished completing, their statistics course. This study is the first to observe and describe students' thinking several months after they completed their course.

While prior theories would have predicted that students would forget some of the information they were taught, those theories could not predict what information would be forgotten and what would be retained. This study helps to close that gap by describing

students' thinking in the context of statistical tests seven months after they completed an introductory level statistics course. Even more so, this study's participants' thinking showed that using the bootstrap dot plot as a basis for thinking about uncertainty and variability may be a more effective approach than using the null model. Furthermore, the participants' thinking showed that they were uncomfortable thinking about statistical testing in anything but the ideal case, implying the need for more complex practice and instruction in the monitoring and evaluating needed to conduct statistical tests.

The brain is a highly complex muscle and students' statistical thinking is an advanced skill that is neither easy to develop nor easy to research. A desirable learning outcome for introductory statistics students is that they learn to think statistically – and continue to do so long after they have left our classrooms. This dissertation scratches the surface of the complexity of students' thinking in statistical tests once they leave the classroom. It calls for new research methods to investigate students' statistical thinking longitudinally and calls for this longitudinal evidence to inform content and pedagogy of introductory statistics courses. Most importantly, by describing what it is that students remember in terms of thought processes, it provides a glimmer of the possibilities for future research on the teaching of statistics.

References

American Psychological Association. (2017). Guidelines for education and training at the doctoral and postdoctoral level in consulting psychology (CP)/ organizational consulting psychology. Retrieved from http://preview.apa.org/about/policy/educationtraining.pdf

Baddeley, A. (2003). Working memory and language: An overview. *Journal of communication disorders, 36*(3), 189-208.

Baddeley, A. D., & Hitch, G. (1993). The recency effect: Implicit learning with explicit retrieval?. *Memory & Cognition*, *21*(2), 146-155.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 353–394). New York: Longman.

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing Critical Thinking. *Journal of Curriculum Studies*, *31*(3), 285–302.

Bandyopadhyay, P. S., & Forster, M. R. (2010). Philosophy of statistics: An introduction. In P. Bandyopadhyay & M. Forster (Eds.) *Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.* Elsevier BV.

Barnard, G. A. (1967). The use of the likelihood function in statistical practice. In L. M. Le Cam, J. Neyman (Eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 27-40.

Baumann, J. F., Seifert-Kessell, N., & Jones, L. A. (1992). Effect of think-aloud instruction on elementary students' comprehension monitoring abilities. *Journal of Reading Behavior, 24*(2), 143-172.

Beckman, M. D. (2015). *Assessment of cognitive transfer outcomes for students of introductory statistics.* Doctoral dissertation, University of Minnesota.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *Challenge of developing statistical literacy, reasoning, and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526-542.

Biehler, R., Frischemeier, D., & Podworny, S. (2015). Preservice teachers' reasoning about uncertainty in the context of randomization tests. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 129-162). Minneapolis, MN: Catalyst Press.

Biehler, R., Kombrink, K., & Schweynoch, S. (2003). MUFFINS – Statistik mit komplexen Datensätzen – Freizeitgestaltung und Mediennutzung von Jugendlichen. *Stochastik in der Schule, 23*(1), 11-25.

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin, 16*(10), 335–338. https://doi.org/10.1037/h0074554

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10*(3), 252–268.

Brown, J. M. (2019). *The extent of quantitative empirical evidence for learning from simulations in statistics courses*. Unpublished Manuscript.

Brown, J. M. (2021). *Student understanding of the hypothetical nature of simulations in introductory statistics.* Doctoral dissertation, University of Minnesota.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. J. (2013). Dynamic Visualizations and the Randomization Test. *Technology Innovations in Statistics Education, 7*(2). http://dx.doi.org/10.5070/T572013889

Carsey, T. M., & Harden, J. J. (2014). *Monte Carlo simulation and resampling methods for social science*. Sage Publications.

Case, C. (2016). *Reasoning about inference using traditional and simulation-based inference models*. Doctoral dissertation, University of Florida.

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.

Çetinkaya-Rundel, M., & Hardin, J. (2021). *Introduction to Modern Statistics*. OpenIntro.

Chance, B., Mendoza, S., & Tintle, N. (2018). Student gains in conceptual understanding in introductory statistics with and without a curriculum focused on simulation-based inference. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward.* Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Chance, B., & Rossman, A. (2006, July). Using simulation to teach and learn statistics. In *Proceedings of the Seventh International Conference on Teaching Statistics* (pp. 1-6). Voorburg, The Netherlands: International Statistical Institute.

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference. *Statistics Education Research Journal, 21*(3), 4. https://doi.org/10.52041/serj.v21i3.6

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education, 24*(3), 114-126.

Chernick, M. R. (2012). Resampling methods. *WIREs Data Mining and Knowledge Discovery*, *2*, 255–262.

Cho, H., Powell, D., Pichon, A., Kuhns, L. M., Garofalo, R., & Schnall, R. (2019). Eye-tracking retrospective think-aloud as a novel approach for a usability evaluation. *International journal of medical informatics, 129*, 366-373.

Clark, M. C. (2010). Narrative learning: Its contours and its possibilities. *New directions for adult and continuing education, 126*(3), 3-11.

Clark, M. C., & Rossiter, M. (2008). Narrative learning in adulthood. *New directions for adult and continuing education, 119*, 61-70.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology innovations in statistics education, 1*.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American mathematical monthly, 104*(9), 801-823.

Cohen, J. (1994). The earth is round (*p*<. 05). *American psychologist, 49*(12), 997-1003.

Cowan, J. (2019). The potential of cognitive think-aloud protocols for educational action-research. *Active Learning in Higher Education, 20*(3), 219-232.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences, 117*(48), 30055-30062.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science, 25*(1), 7-29.

de Finetti, B. (1980). Foresight: Its logical laws, its subjective sources (H.E. Kyburg Jr., trans.). In H. E. Kyburg, Jr. and H. E. Smokler (Eds.), *Studies in Subjective Probability*. John Wiley and Sons. (Original work published 1937)

DeLiema, D., Hufnagle, A. S., Rao, V. N. V., Baker, J., Valerie, J., & Kim, J. (2023). Methodological considerations and innovations at the intersection of video-based research traditions. *International Journal of Research and Method in Education, 46*(1), 19-36. https://doi.org/10.1080/1743727X.2021.2011196

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 79-95). Dordrecth: Springer. https://doi.org/10.1007/1-4020-2278-6_4

delMas, R. (2021). *randomizeIt: Randomization Methods for Bootstrapping and Hypothesis Testing*. R package version 0.1.0.

delMas, R. C., Garfield, J. B., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3). https://doi.org/10.1080/10691898.1999.12131279

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing Students' Conceptual Understanding after a First Course in Statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, *28*(3), 425-474.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational psychology review*, *20*(4), 391-409.

Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 705-725). Cambridge University Press: New York.

Durning, S. J., Artino Jr., A. R., Beckman, T. J., Graner, J., Van Der Vleuten, C., Holmboe, E., & Schuwirth, L. (2013). Does the think-aloud protocol reflect thinking? Exploring functional neuroimaging differences with thinking (answering multiple choice questions) versus thinking aloud. *Medical teacher, 35*(9), 720-726.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press (1st Ed.).

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*(3), 178-186.

Ernst, M. D. (2004). Permutation methods: A Basis for Exact Inference. *Statistical Science, 19*, 676-685.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. *Theory and Psychology, 5*(1), 75–98.

Fan, M., Lin, J., Chung, C., & Truong, K. N. (2019). Concurrent think-aloud verbalizations and usability problems. *ACM Transactions on Computer-Human Interaction*, 26(5), 1-35.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd: London.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd: London.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Hafner Publishing Co.

Fisher, W. R. (1984). Narration as a Human Communication Paradigm: The Case of Public Moral Argument. *Communication Monographs, 51*, 1–18.

Frischemeier, D. & Biehler, R. (2013). Design and exploratory evaluation of a learning trajectory leading to do randomization tests facilitated by TinkerPlots. In B. Ubuz, C. Haser, & M. A. Mariotti (Eds.), *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education* (pp. 799–809).

Fry, E. B. (2017). *Introductory statistics students' conceptual understanding of study design and conclusions.* Doctoral dissertation, University of Minnesota.

GAISE (2016). *Guidelines for assessment and instruction in statistics education*. College report. Alexandria, VA: American Statistical Association.

Garfield, J., delMas, R.C., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM, 44*(7), 883-898.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Ed.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606.

Glaser, B. G., & Strauss, A. L. (1967). Discovery of grounded theory: Strategies for qualitative research. Taylor & Francis. (Original work published in 1967). https://doi.org/10.4324/9780203793206

Glencross, M. J. (1988). A Practical Approach to the Central Limit Theorem. In *Proceedings of the Second International Conference on Teaching Statistics (ICOTS),* Victoria, B.C. (pp 91-95). The Organizing Committee for the Second ICOTS.

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.

Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative comprehension. *Psychological Review, 101*, 371-395.

Hacking, I. (1965). *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hájek, A. (2019). Interpretations of probability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 Edition).

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

Hempel, C. G. (1945). Studies in the Logic of Confirmation. *Mind, 54*(213), 1-26.

Henderson, L. (2020). The problem of induction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University.

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and understanding in introductory statistics under consensus and simulation-based curricula. *Statistics Education Research Journal, 17*(1), 103-120.

Hogben, L. (1957). *Statistical theory*. London: Allen & Unwin.

Howard, B. C., McGee, S., Shia, R., & Hong, N. S. (2000, April). Metacognitive self-regulation and problem-solving: Expanding the theory base though factor analysis [Paper Presentation]. Annual Meeting of the American Educational Research

Association, New Orleans, Louisiana. Available at https://files.eric.ed.gov/fulltext/ED470973.pdf.

Hyrskykari, A., Ovaska, S., Majaranta, P., Räihä, K. J., & Lehtinen, M. (2008). Gaze path stimulation in retrospective think-aloud. *Journal of Eye Movement Research, 2*(4), 5: 1-18.

Jahn, M., Herman, D., & Ryan, M.-L. (2010). *Routledge Encyclopedia of Narrative Theory*. Taylor and Francis.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press (3rd ed.).

Justice, N., Le, L., Sabbag, A., Fry, E., Ziegler, L., & Garfield, J. (2020). The CATALST Curriculum: A Story of Change. *Journal of Statistics Education, 28*(2), 175-186.

Justice, N., Zieffler, A., Huberty, M. D., & delMas, R. C. (2018). Every rose has its thorn: secondary teachers' reasoning about statistical models. *ZDM, 50*(7), 1253-1265.

Kaplan, A., Lichtinger, E., & Gorodetsky, M. (2009). Achievement goal orientations and self-regulation in writing: An integrative perspective. *Journal of Educational Psychology, 101*(1), 51-69.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan and Company, limited.

Kitchner, K. S. (1983). Cognition, metacognition, and epistemic cognition. *Human development, 26*(4), 222-232.

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning, 12*, 217-230.

Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic data exploration*. Key Curriculum Press: Emeryville, CA.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178-206.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology*, *22*(1), 67-90.

Le, L. (2017). *Assessing the Development of Students' Statistical Thinking: An Exploratory Study*. Doctoral dissertation, University of Minnesota.

Lenhard, J. (2006). Models and Statistical Inference: The controversy between Fisher and Neyman-Pearson. *British Journal of the Philosophy of Science, 57*, 69-91.

Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2021), *Statistics*: *Unlocking the Power of Data*, Hoboken, NJ: Wiley.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, *25*(12), 720–725. https://doi.org/10.3928/0048-5713-19951201-07

Maehr, M. L. (1984). Meaning and motivation: Toward a theory of personal investment. In C. Ames, & R. Ames (Eds.), *Research on motivation in education*, Vol. 1, pp. 115–144. New York: Academic.

Maher, C.A., Sigley R. (2014) Task-Based Interviews in Mathematics Education. In: Lerman S. (eds) *Encyclopedia of Mathematics Education*. Springer, Dordrecht.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82-105.

Maxara, C., & Biehler, R. (2007). Constructing stochastic simulations with a computer tool – students' competencies and difficulties. In D. Pitta & P. G. Philippou (Eds.),

Proceedings of the Fifth Conference of the European Society for Research in Mathematics Education (p. 762-771). Lamaca, Cyprus.

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational psychologist, 35*(1), 13-23.

Mendoza, S., & Roy, S. (2018). Assessing retention of statistical concepts after completing a post-secondary introductory statistics course. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward.* Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Merriam, S. B. (1988). *Case study research in education: A qualitative approach*. Jossey-Bass.

Meyer, B. J. F., & Freedle, R. O. (1984). Effects of discourse type on recall. *American Educational Research Journal, 21*, 121-143.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14-20.

Moore, D. S. (1990). Uncertainty. In L.A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-138). Washington, D. C.: National Academy Press

Moore, D. S. (1992). Teaching Statistics as a Respectable Subject. In F. Gordon & S. Gordon (Eds.), *Statistics for the Twenty-First Century.* MAA Notes: Mathematical Association of America.

Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). *The basic practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

Morgan, K. L., Lock, R. H., Lock, P. F., Lock, E. F., & Lock, D. F. (2014, July). StatKey: Online tools for bootstrap intervals and randomization tests. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education*. Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Morrison. D. E., & Henkel, R. E. (1970). *The significance test controversy*. Chicago: Aldine.

Moshman, D., & Tarricone, P. (2016). Logical and causal reasoning. In J. A. Greene, W. A. Sandoval, I. Bråten (Eds.) *Handbook of epistemic cognition* (pp. 54-67). Routledge.

Muijs, D. and Bokhove, C. (2020). *Metacognition and Self-Regulation: Evidence Review*. London: Education Endowment Foundation.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*(767), 333-380.

Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability* (2nd ed.). Washington: The Graduate School U.S. Department of Agriculture.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 175-240.

Nicholas, M. (2018). Affordances of using Multiple Videoed Events to construct a rich understanding of adult–child book readings. *International Journal of Research & Method in Education, 41*(2), 125-141.

Nickerson, R. S. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods, 5*(2), 241-301.

Noll, J., Clement, K., Dolor, J., Kirin, D., & Petersen, M. (2018a). Students' use of narrative when constructing statistical models in TinkerPlots. *ZDM, 50*(7), 1267-1280.

Noll, J., Clement, K., Dolor, J., & Peterson, M. (2018b). Students' statistical modeling activities using TinkerPlots. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward.* Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Noll, J., Gebresenbet, M., & Glover, E. D. (2016). A modeling and simulation approach to informal inference: Successes and challenges. In D. Ben-Zvi & K. Makar (Eds.), *The teaching and learning of statistics: International perspectives* (pp.139-150). New York: Springer

Noll, J. & Kirin, D. (2016). Student approaches to constructing statistical models using TinkerPlots™. *Technology Innovations in Statistics Education, 9*(1).

Noll, J. & Kirin, D. (2017). TinkerPlots model construction approaches for comparing two groups: Student perspectives. *Statistics Education Research Journal, 16*(2), 213-243.

Noll, J., Kirin, D., Clement, K., & Dolor, J. (2023). Revealing students' stories as they construct and use a statistical model in TinkerPlots to conduct a randomization test for comparing two groups. *Mathematical Thinking and Learning, 25*(1), 44-63. https://doi.org/10.1080/10986065.2021.1922858

Norton, J. (2005). A little survey on induction. In P. Achinstein (Ed.), *Scientific evidence: Philosophical theories and applications* (pp. 9-34). Baltimore: John Hopkins University Press.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50*(302), 157-175.

Polkinghorne, D. E. (1995). Narrative configuration in qualitative analysis. *International journal of qualitative studies in education, 8*(1), 5-23.

Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science, 10*(37), 25-42.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao, C. R. (1992). RA Fisher: The founder of modern statistics. *Statistical Science, 7*(1), 34-48.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68.

Ricoeur, P. (2005). *Hermeneutics and the human sciences* (17th edn.) (J. B. Thompson, Trans., Ed.). Cambridge, MA: Cambridge University Press.

Robinson, J. A., & Hawpe, L. (1986). Narrative thinking as a heuristic process. In T. R. Sarbin (Ed.), *Narrative psychology: The storied nature of human conduct* (pp. 111–125). Praeger Publishers/Greenwood Publishing Group.

Romeijn, J. W. (2017). Philosophy of Statistics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 Edition).

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*(4), 211-221.

Rossman, A. Chance, B., & Lock, R.H. (2001). *Workshop Statistics: Discovery with Data.* New York: Key College Publishing.

Rowe, J. P., Mcquiggan, S. W., Mott, B. W., & Lester, J. C. (2007). Motivation in narrative-centered learning environments. In *Supplementary Proceedings of the 13th International Conference on Artificial Intelligence in Education (AIED).* Marina Del Rey, California, USA.

Roy, S., & Mcdonnel, T. (2018). Assessing simulation-based inference in secondary schools. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward.* Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 335-391). Hillsdale, NJ: Erlbaum.

Sabbag, A. G. (2016). *Examining the relationship between statistical literacy and statistical reasoning.* Doctoral dissertation, University of Minnesota.

Sabbag, A. G., Garfield, J., & Zieffler, A. (2015). Quality Assessments in Statistics Education: A Focus on the GOALS Instrument. In M.A. Sorto (Ed.) *Advances in Statistics Education: Developments, Experiences, and Assessments*. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), Rio de Janeiro, Brazil.

Schank, R. C. (2000). *Tell me a story: Narrative and intelligence*. Evanston, IL: Northwestern University Press.

Schoenfeld, A. H. (1985). Making sense of "out loud" problem-solving protocols. *The Journal of Mathematical Behavior, 4*(2), 171-191.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in science education, 36*(1), 111-139.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational psychology review, 7*(4), 351-371.

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational psychology review, 20*(4), 463-467.

Simon, M. A. (2019). Analyzing Qualitative Data in Mathematics Education. In K. R. Leatham (Ed), *Designing, Conducting, and Publishing Quality Research in*

Mathematics Education. Research in Mathematics Education. Springer, Cham. https://doi.org/10.1007/978-3-030-23505-5_8

Smagorinsky, P. (1998). Thinking and speech and protocol analysis. *Mind, culture, and activity, 5*(3), 157-177.

Snee, R. D. (1993). What's missing in statistical education? *The American Statistician, 47*(2), 149-154.

Son, J. Y., Blake, A. B., Fries, L., & Stigler, J. W. (2021). Modeling first: Applying learning science to the teaching of introductory statistics. *Journal of Statistics and Data Science Education, 29*(1), 4-21.

Sprenger, J. (2011). Hypothetico-deductive confirmation. *Philosophy Compass, 6*(7), 497-508.

Stanley, J. C. (1966). The influence of Fisher's "The Design of Experiments" on educational research thirty years later. *American Educational Research Journal, 3*(3), 223-229.

Stigler, S. M. (2016). *The seven pillars of statistical wisdom.* Harvard University Press.

Talbott, W. (2016). Bayesian Epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 Edition).

Tintle, N. L., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2020). *Introduction to statistical investigations*. Wiley & Sons.

Tintle, N. L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., … VanderStoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education*. Proceedings of the Ninth International

Conference on Teaching Statistics, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Tintle, N. L., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. Statistics Education Research Journal, 11(1), 21-40.

Tintle, N., VanderStoep, J., Holmes, V. L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1).

Tobías-Lara, M. G., & Gómez-Blancarte, A. L. (2019). Assessment of informal and formal inferential reasoning: A critical research review. *Statistics Education Research Journal, 18*(1), 8-25.

Vallecillos, A. & Batanero, C. (1996). Conditional probability and the level of significance in tests of hypotheses. In L. Puig & A. Gutiérrez (Eds.), *Proceedings of the 20th conference of the International Group for the Psychology of Mathematics Education*, University of Valencia, Valencia, Spain.

VanderStoep, J. L., Couch, O., & Lenderink, C. (2018). Assessing the association between quantitative maturity and student performance in an introductory statistics class: Simulation-based vs non simulation-based. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward.* Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

von Mises, R. (1957). *Probability, Statistics, and Truth* (J. Neyman, D. Scholl, and E. Rabinowitsch, trans.). New York: Macmillan. (Original work published 1939)

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics, 10*(4), 299-326.

Wang, Y., Zhang, D., Du, G., Du, R., Zhao, J., Jin, Y., ... & Hu, Y. (2020). Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet*.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond p< 0.05. *The American Statistician, 73*(S1), 1-19.

Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., Almeqdad, Q., & Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning, 4*(1), 63-85.

Wiggins, S. (2016). *Discursive psychology: Theory, method and applications*. Sage.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Ziedner (Eds.), *Handbook of self-regulation* (pp. 531–566). San Diego, CA: Academic

Winters, F. I., Greene, J. A., & Costich, C. M. (2008). Self-regulation of learning within computer-based learning environments: A critical analysis. *Educational Psychology Review, 40,* 429–444.

World Health Organization (WHO). (2020, March 11). *WHO Director-General's opening remarks at the media briefing on COVID-19*. Retrieved from:

https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Zieffler, A., & Catalysts for Change. (2021). *Statistical Thinking: A simulation approach to uncertainty* (4.3rd ed.). Minneapolis, MN: Catalyst Press. http://zief0002.github.io/statistical-thinking/

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40–58.

Table 1

Students' scores on Tests of Significance items in studies comparing curricula by assessment

Source	Traditional curriculum				ISI curriculum			
	Students	Pretest	Posttest	Difference	Students	Pretest	Posttest	Difference
				CAOS Tests of Significance items				
Tintle et al. (2011)	195	48.8%	61.5%	12.7%	202	50.0%	69.8%	19.8%
Tintle et al. (2012)	78	51.5%	67.3%	15.8%	76	51.5%	71.3%	19.8%
Tintle et al. (2014)	94	50.0%	60.6%	10.6%	155	46.1%	70.0%	23.9%
				ISI Tests of Significance items				
Chance et al. (2016)	~60*	50.4%	55.8%	5.4%	~1050*	57.7%	68.9%	11.2%
Mendoza & Roy (2018)	284	58.0%	60.9%	2.9%	197	58.0%	69.5%	11.5%
Roy & Mcdonnel (2018)	435	-	-	6.3%	196	-	-	14.6%
VanderStoep et al. (2018)	601^	40.3%	48.8%	8.5%	886^	39.1%	58.3%	19.2%

*Exact number of student respondents per group was not reported, and instead are estimated based on a total of 1116 students across 34 simulation-based sections and 2 traditional sections.
^Results by topic only reported for students scoring less than 40% overall.

Table 2

Frischemeier and Biehler's (2013) randomization test plan with examples

No.	Step	Example solution to the ESP task (Rossman et al., 2001)	Expected solution to the Muffins task (Biehler et al., 2003)
	Observation		
1	Which difference do you observe between the means of the two groups in the dataset?	Number of correct answers = 20	Mean of Time_Reading of boys = 2.685 Mean of Time_Reading of girls = 3.503 Difference = 0.818
	Hypothesis		
2	As said in the task, the difference of the means of the two groups could have occurred at random. Generate an adequate Null Hypothesis for your investigation.	The person does not have extrasensory perception (ESP). He/She guesses with a success rate p = 0.25.	The difference of the means of Time_Reading of boys and girls has occurred at random.
	Simulation of H0		
3	How can you investigate the null hypothesis with a simulation? Explain your procedure.	Drawing 40 times with replacement from an urn which is filled with 4 balls: 1 ball is labeled "right" and 3 balls are labeled "false".	Place the 533 cases of Time_Reading in urn1. Construct urn2 with 232 balls labeled "male" and 301 balls labeled "female". Draw 533 times without replacement.

4	**Test Statistic** Define the test statistic.	X = number of correct predictions	X = mean of group 1 minus mean of group 2
5	***p*-value** Calculate the *p*-value	P(X>20) = 0.0004	P(X>0.818) = 0.0006
6	**Conclusions** Which conclusions can you make regarding your null hypothesis?	The *p*-value is very small, so we have strong evidence against our null hypothesis. We assume that the fortune teller has not guessed. Another possibility is: he could have guessed but that would have been very unlikely.	The *p*-value is very small, so we have strong evidence against our null hypothesis. Another possibility is: the difference occurred at random, but that is very unlikely.

Table 3

Examples of empirical traces of planning, monitoring, and evaluating

Task	Planning	Monitoring	Evaluating
Concept Mapping Task	Chau planned to compute a *p*-value to answer the research question. "To answer the question, we use only the *p*-value" (Appendix P06-A, 06:54 – 06:58).		Tal explained that finding non-sensible results would prompt an evaluation of the steps Tal had conducted, and prompt quality checks on the analysis process. "If my conclusion doesn't make sense, it may be my own error. So I have to figure that out also" (Appendix P02-A, 03:28 – 03:34).
Statistical Testing Task	Jaci made a plan to compute and interpret a *p*-value to determine whether a difference between group means was significant. "I mean I guess looking at the *p*-value would be the number one tell. And I think that a low *p*-value means that it's pretty, it is pretty likely that there is a significant difference" (Appendix P04-B1, 12:42 – 13:03).	Chau monitored which statistics they had generated while completing the VSE problem. "So I have all the data, median, mean, and then quartile, for both of the groups, standard deviation" (Appendix P06-B1, 07:50 – 08:00).	While completing the VSE problem in R, and after having drawn a conclusion based on the confidence interval and *p*-value, Jaci evaluated that they may need to create a bootstrap dot plot. "Should I be trying to make a graph of this?" (Appendix P04-B1, 08:28 – 08:33).

Statistical Testing Interview	Kei planned to inspect *p*-values when using StatKey, primarily due to feeling overwhelmed by the manner in which StatKey presented information. "I feel like there's so much stuff to look at in that [StatKey] graph that I just shut down … but here [in R] it's very clear" (Appendix P03-C, 06:25 – 06:40).	Ade monitored the results of the comparison of sample means by examining the *p*-value. "well it has a low *p*-value so I feel like this is a good sample. And then, just the means alone, 81 and 67 are pretty different" (Appendix P05-C, 17:36 – 17:46).	Tal explained that if there is an estimated difference between groups, they would next plan to evaluate the sensibility of the results. "One of the first things that I do is, especially when you have a difference in averages, I'm going to see if that means if it's really different" (Appendix P02-C, 15:15 – 16:35).
Video-Cued Interview	Kei planned to generate a graph displaying the sample data for each group. "I need to see each individual graph … and that's how I was taught, I don't know any other way" (Appendix P03-D, 12:30 – 12:36).	Jaci monitored the tails of the null model in StatKey to ensure that the their interpretation of the confidence interval was correct. "I think I was looking essentially to see like, if the, like what the different parameters were, to see whether or not they were truly fairly centered around zero, or if there was maybe a difference on either side" (Appendix P04-D, 08:02 – 08:25).	

Table 4

Kei's step-by-step plan for statistical testing

No.	Step	Explanation
1	Import necessary libraries and packages	(R Only)
2	Import the data	
3	Check the data	(using the names(), head(), and tail() functions in R)
4	Compute the mean for Group 1	"Well, I'd want the average salary of both [groups]" (Appendix P03-B1, 04:15 – 04:18) "I would choose one of the salaries, errr, one of the, choose one of the majors first." (Appendix P03-B1, 08:10 – 09:00) "I wanted to look at each individual major and salary, and examine that." (Appendix P03-D, 10:37 – 10:51)
5	Create a graph for Group 1	"I would choose one of the salaries, errr, one of the, choose one of the majors first, and then I'd want to do like a little graph thing." (Appendix P03-B1, 08:10 – 09:00) "I need to see each individual graph" (Appendix P03-D, 12:30 – 12:36) "So then I would want to create … a graph … for each, I'm interested to see like what it looks like, for like the range" (Appendix P03-B1, 13:24 – 13:38)
6	Compute the mean for Group 2	

7	Create a graph for Group 2	"When I'm plotting it out, I'm thoroughly examining and thinking about each individual group, I'm not just jumping into the final comparison. I'm like, okay, what's going on with this group, [then] what's going on with this [other] group, and then looking at the *p*-value." (Appendix P03-D, 14:26 – 14:48)
8	Compute the difference in means	"cuz like, [the research question is] 'Is there a difference in the average salary', so I'd want to find the average." (Appendix P03-B1, 16:00 – 16:38)
9	Compute a *p*-value OR Compute a 95% confidence interval estimate	"When I'm plotting it out, I'm thoroughly examining and thinking about each individual group, I'm not just jumping into the final comparison. I'm like, okay, what's going on with this group, [then] what's going on with this [other] group, and then looking at the *p*-value." (Appendix P03-D, 14:26 – 14:48) "I don't even look at the *p*-value for these [outputs in R] but I look at the *p*-values for the graphs [outputs in StatKey] … I feel like I can just think about it [based on the confidence interval]." (Appendix P03-C, 05:53 – 06:19)

Figure 1

Excerpt of the VSE Task used by Biehler et al. (2015)

In the dataset you can see the monthly salaries of 861 women and men of the year 2006. The display suggests that women are way behind men concerning their salary. Someone argues against the result of the group comparison between women and men that only 861 employees were asked. Therefore, the differences could have emerged due to the selection of our sample.

YOUR TASK: Now check if there is evidence against the assumption that there is no difference between women and mean in the population with regard to their average salary. (This would mean that we can expect similar differences for all employees.)

Figure 2

The Dolphin Therapy Task used by Noll and Kirin (2017)

Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the animal care program) did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subject's level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed substantial improvement (reducing their level of depression) by the end of the study (Antonioli and Reveley, 2005). *Research Question: Is swimming with dolphins therapeutic for patients suffering from clinical depression?* The researchers found that 10 of 15 subjects in the dolphin therapy group showed substantial improvement, compared to 3 of 15 subjects in the control group.

The above descriptive analysis tells us what we have learned about the 30 subjects in the study. But can we make any inferences beyond what happened in this study? Does the higher improvement rate in the dolphin group provide convincing evidence that the dolphin therapy is effective? Is it possible that there is no difference between the two treatments and that the difference observed could have arisen just from the random nature of putting the 30 subjects into groups (i.e., the luck of the draw)? We can't expect the random assignment to always create perfectly equal groups, but is it reasonable to believe the random assignment alone could have led to this large of a difference?

The key statistical question is: If there really is no difference between the therapeutic and control conditions in their effects of improvement, how unlikely is it to see a result as extreme or more extreme than the one you observed in the data just because of the random assignment process alone?

Figure 3

The Facebook Task used by Noll and Kirin (2016)

Facebook is a social networking Web site. One piece of data that members of Facebook often report is their relationship status: single, in a relationship, married, it's complicated, etc. With the help of Lee Byron of Facebook, David McCandless - a London-based author, writer, and designer - examined changes in peoples' relationship status, in particular, breakups. A plot of the results showed that there were repeated peaks on Mondays, a day that seems to be of higher risk for breakups.

Consider a random sample of 50 breakups reported on Facebook within the last year. Of these 50, 20% occurred on Monday. Explain how you could determine whether this result would be surprising if there really is no difference in the chance for relationship break-ups among the seven days. *(Be sure to give enough detail that someone else could easily follow your explanation.)*

Figure 4

The Music Note Task from the Models of Statistical Thinking (MOST) assessment (Garfield et al., 2012)

Some people who have a good ear for music can identify the notes they hear when music is played. One note identification test consists of a music teacher choosing one of the seven notes (A, B, C, D, E, F, or G) at random and playing it on a piano. The student is standing in the room facing away from the piano so that they cannot see which note the teacher plays on the piano. The note identification test has the music student identify 10 such notes.

This note identification test was given to a young music student to determine whether or not the student has this ability. The student correctly identifies 7 notes out of the 10 that were played. Explain how you would use what you learned in this class to determine how surprising this result is and whether it is strong evidence that the student has the musical ability to accurately identify notes? *(Be sure to give enough detail that someone else could easily follow your explanation.)*

Figure 5

The NFL Task used by Noll et al. (2018b)

The National (American) Football League (NFL) uses an overtime period to determine a winner for games that are tied at the end of regulation time. Between 1974 and 2009, the overtime period started with a coin flip to determine which team gets the ball first in overtime, and then the team that scores first wins. Data from the 1974 through 2009 seasons show that the coin flip winner won 240 out of the 428 (56%) games where a winner was determined in overtime. Research Question: Is there an advantage to the team that wins the coin flip?

Figure 6

Framework for randomization testing proposed by Biehler et al. (2015)

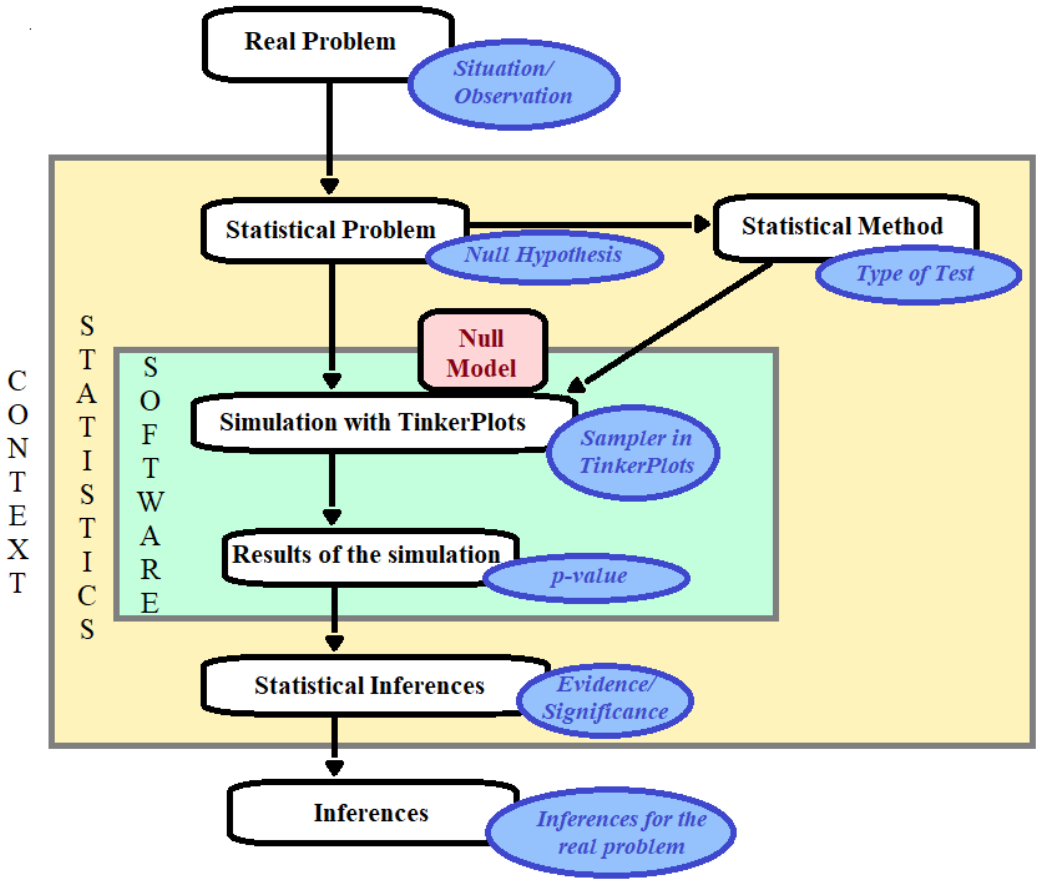


Figure 7

Summary of the study design, tasks, and data artifacts

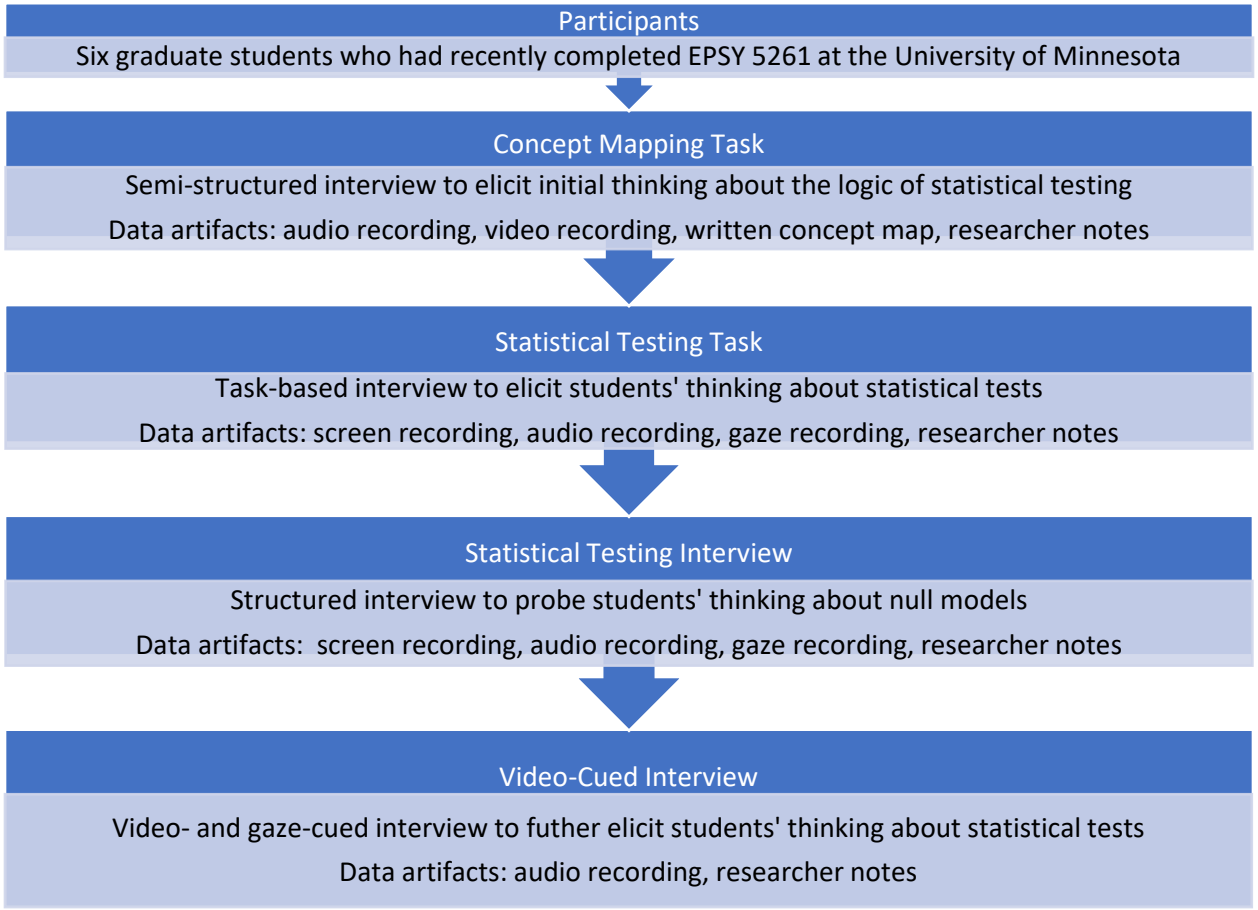


Figure 8

Jaci's concept map for the logic of a statistical test

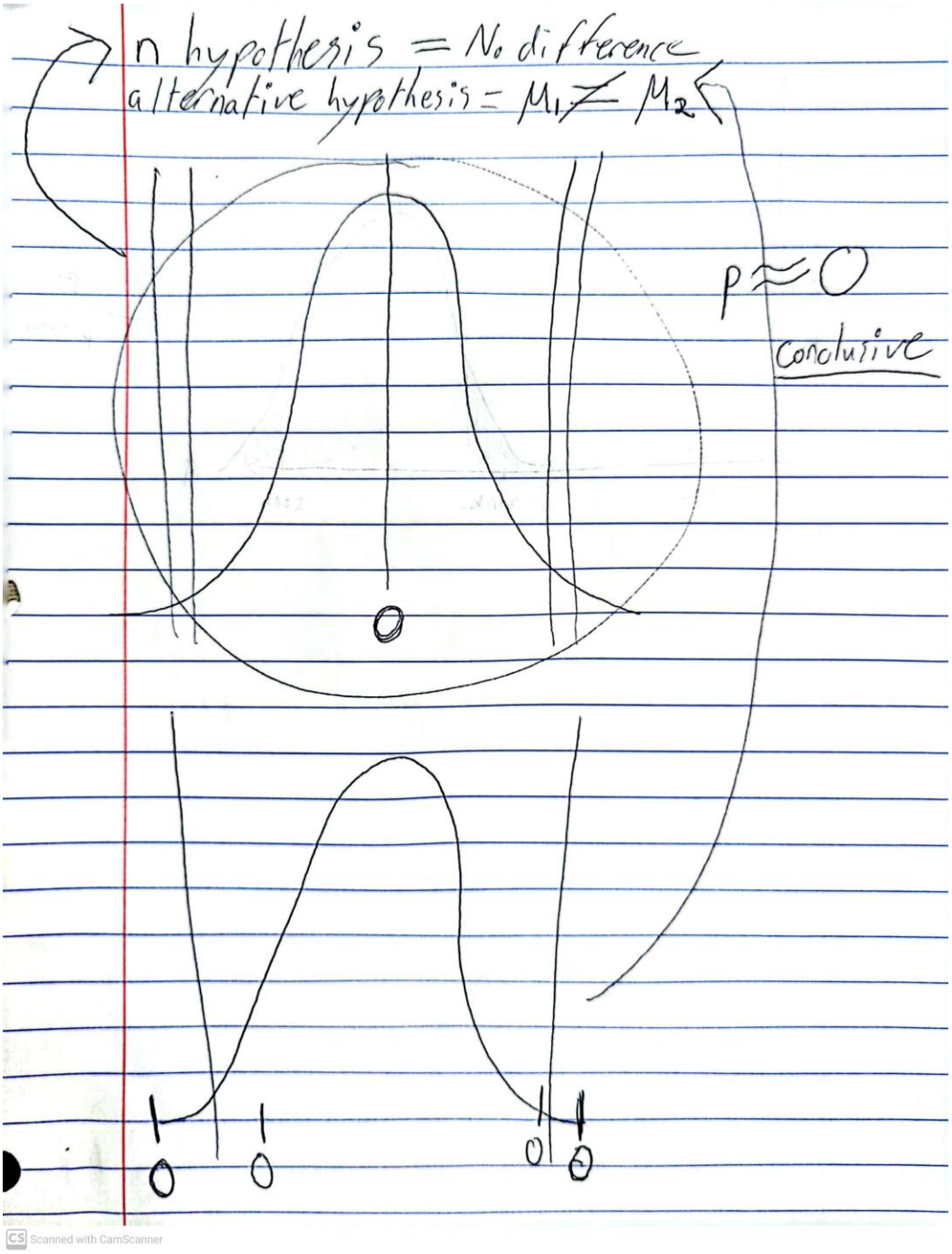


Figure 9

Screen Shot from the Video-Cued Interview (Appendix P04-D, 16:27) in which Jaci is commenting on their process for comparing the observed sample distribution (top right of the screen) to the distribution for a single simulated trial (lower right of the screen)

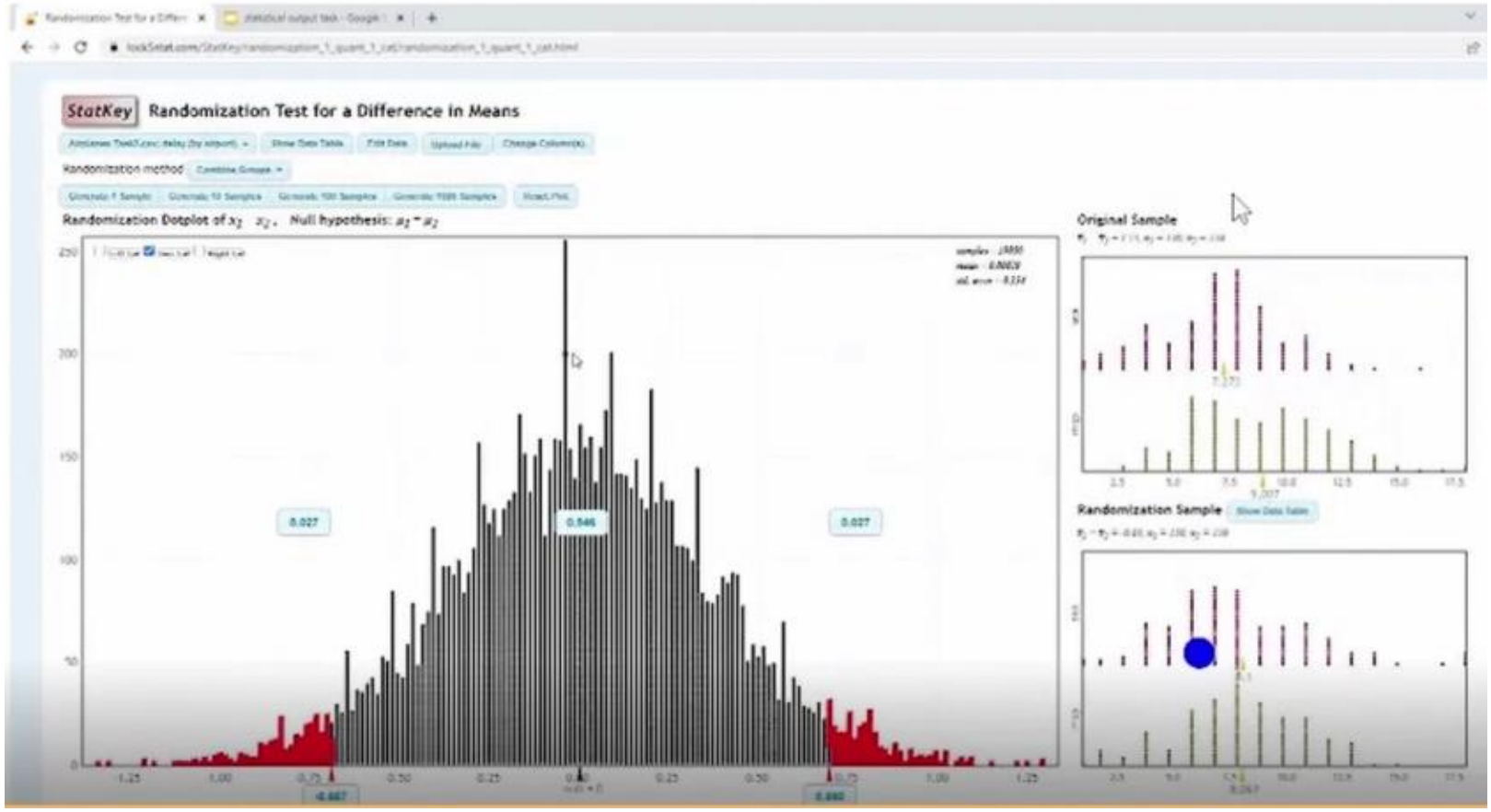


Figure 10

Screenshot of Jaci completing the Airplane Delays Task as part of the Statistical Testing Task (Appendix P04-B2, 03:02), while looking at the mode of the randomization dot plot (green dot), and interpreting this as the most likely outcome

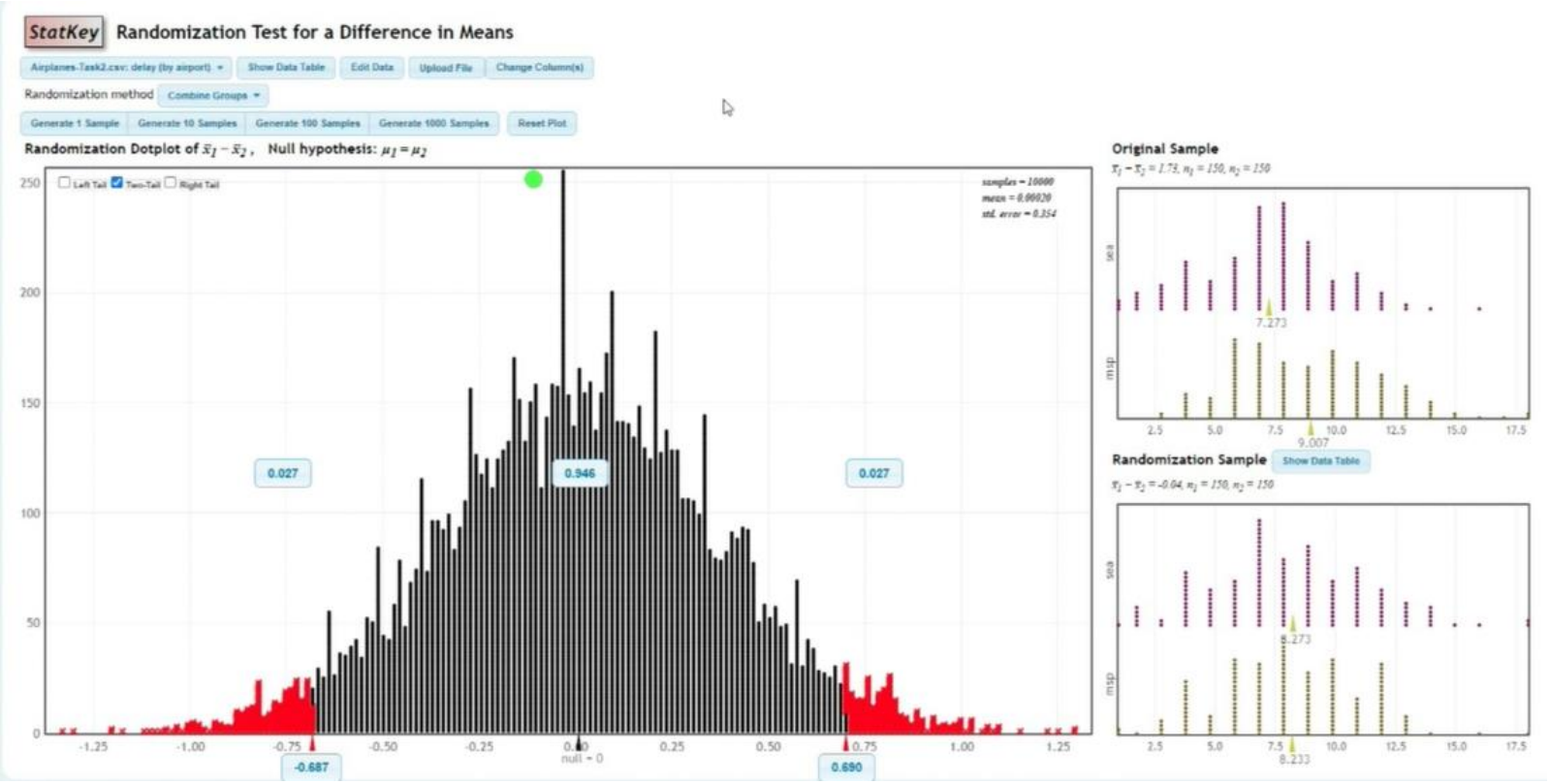


Figure 11

Kei's concept map for the logic of a statistical test

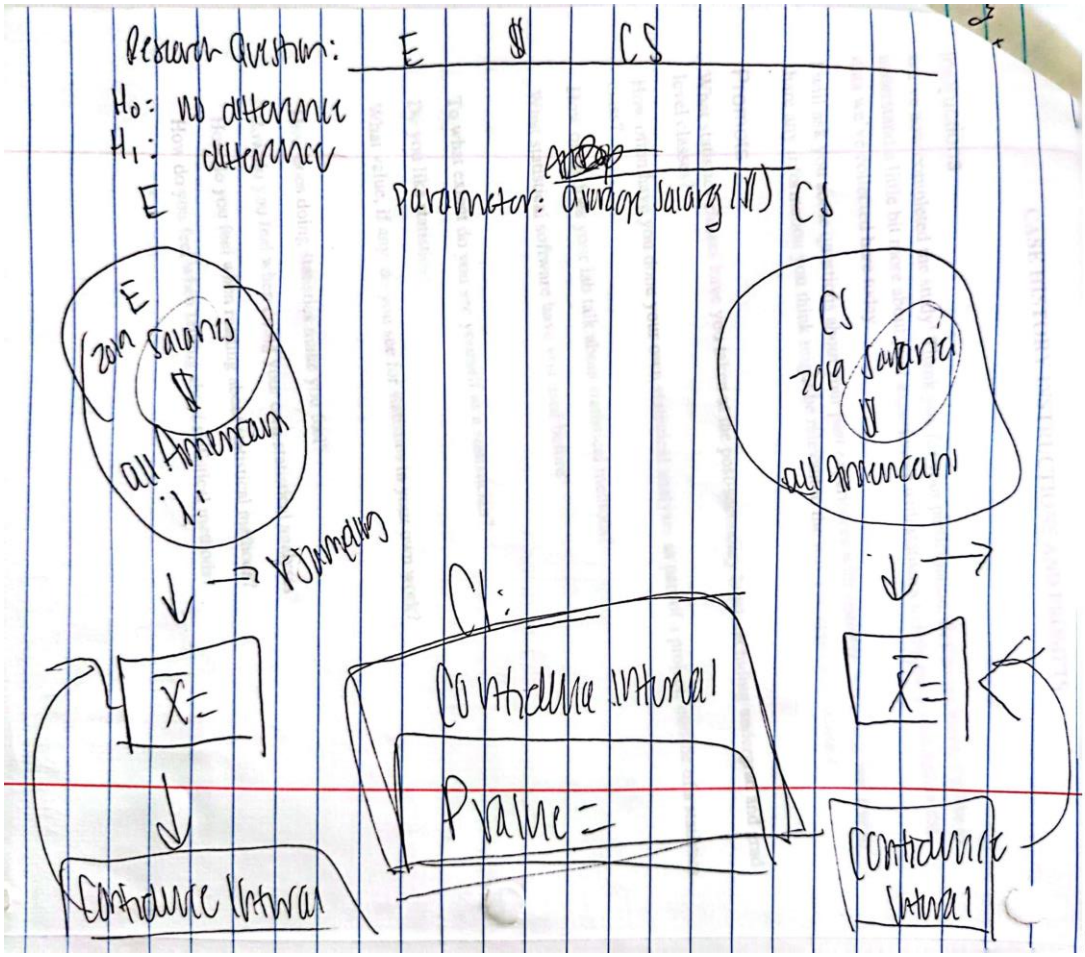


Figure 12

Screenshot of Kei answering the question 'Is average commute time in Atlanta and St Louis the same?' as part of the Statistical Testing Interview and looking at the center of the randomization dot plot (green dot, Appendix P03-C, 01:08)

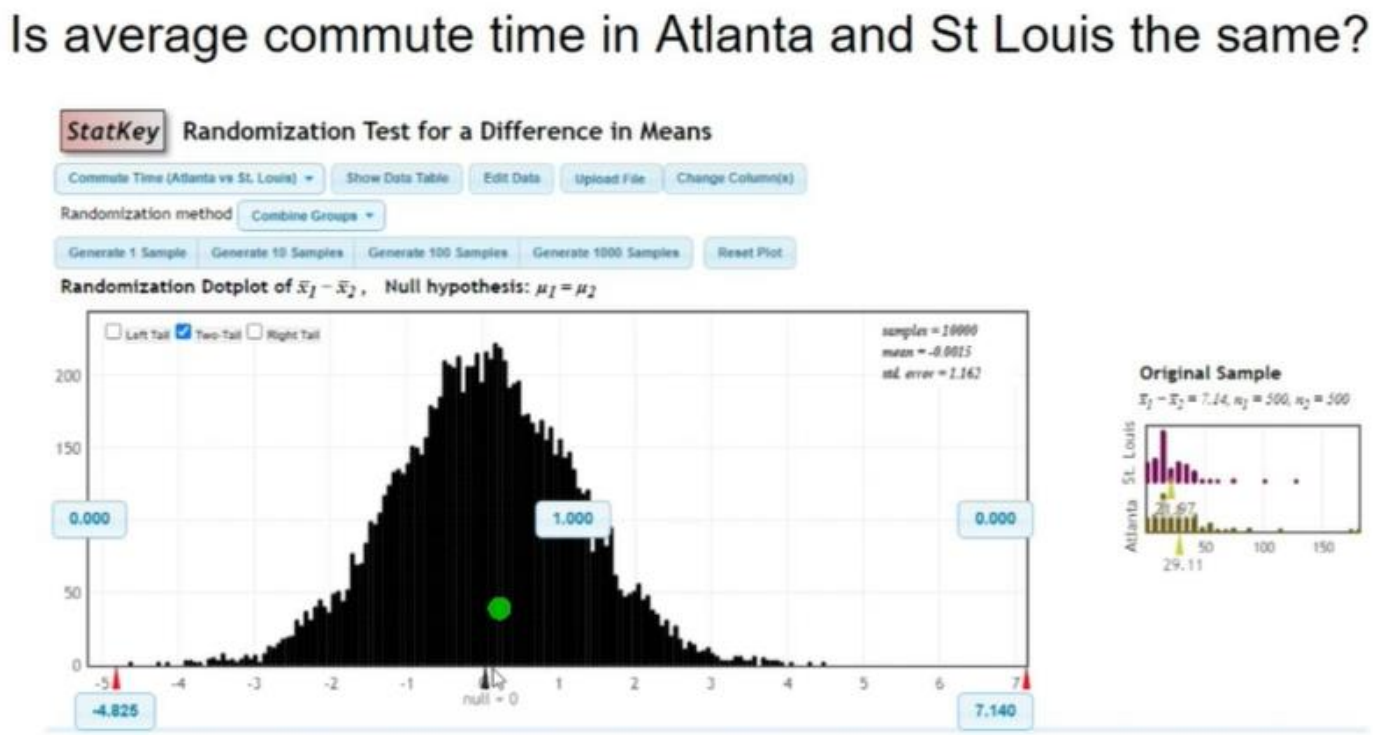


Figure 13

Screenshot of Kei answering the question 'Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?' as part of the Statistical Testing Interview and looking first at the center of the randomization dot plot (green dot, Appendix P03-C, 05:05)

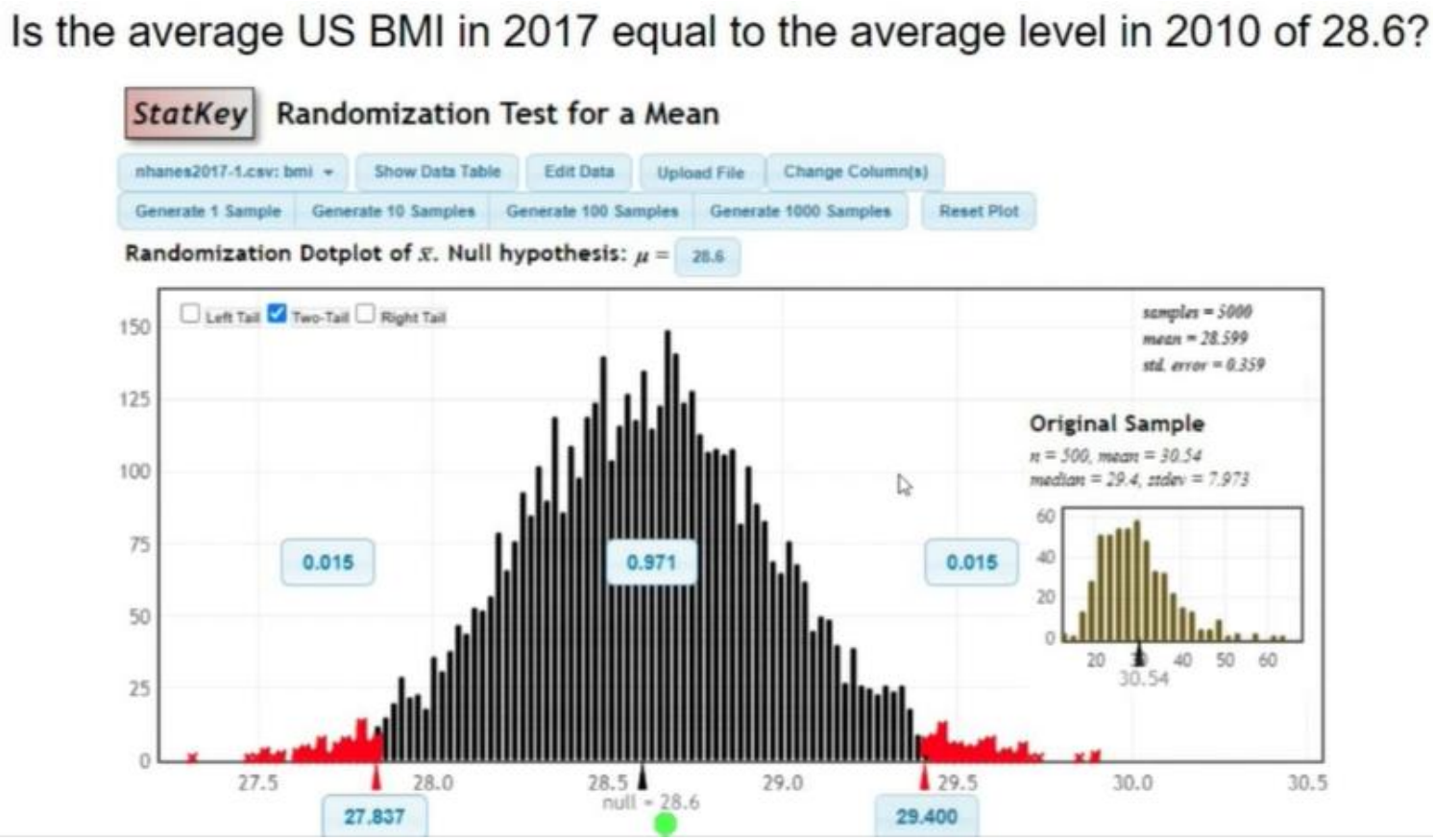


Figure 14

Screenshot of Kei answering the question 'Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?' as part of the Statistical Testing Interview and looking at the p-value (green dot, Appendix P03-C, 05:07), after first looking at the center of the randomization dot plot

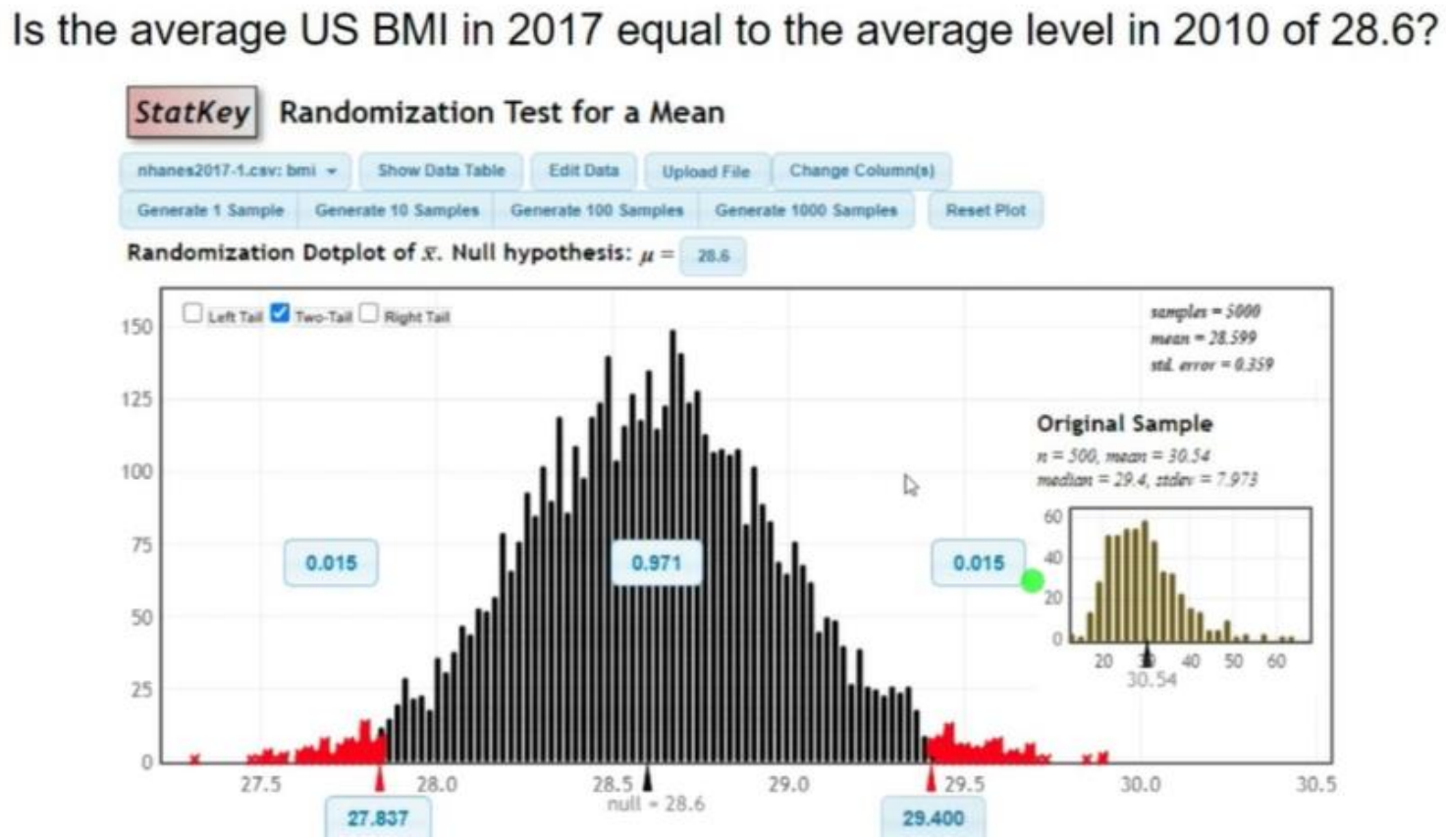


Figure 15

Chau's concept map for the logic of a statistical test

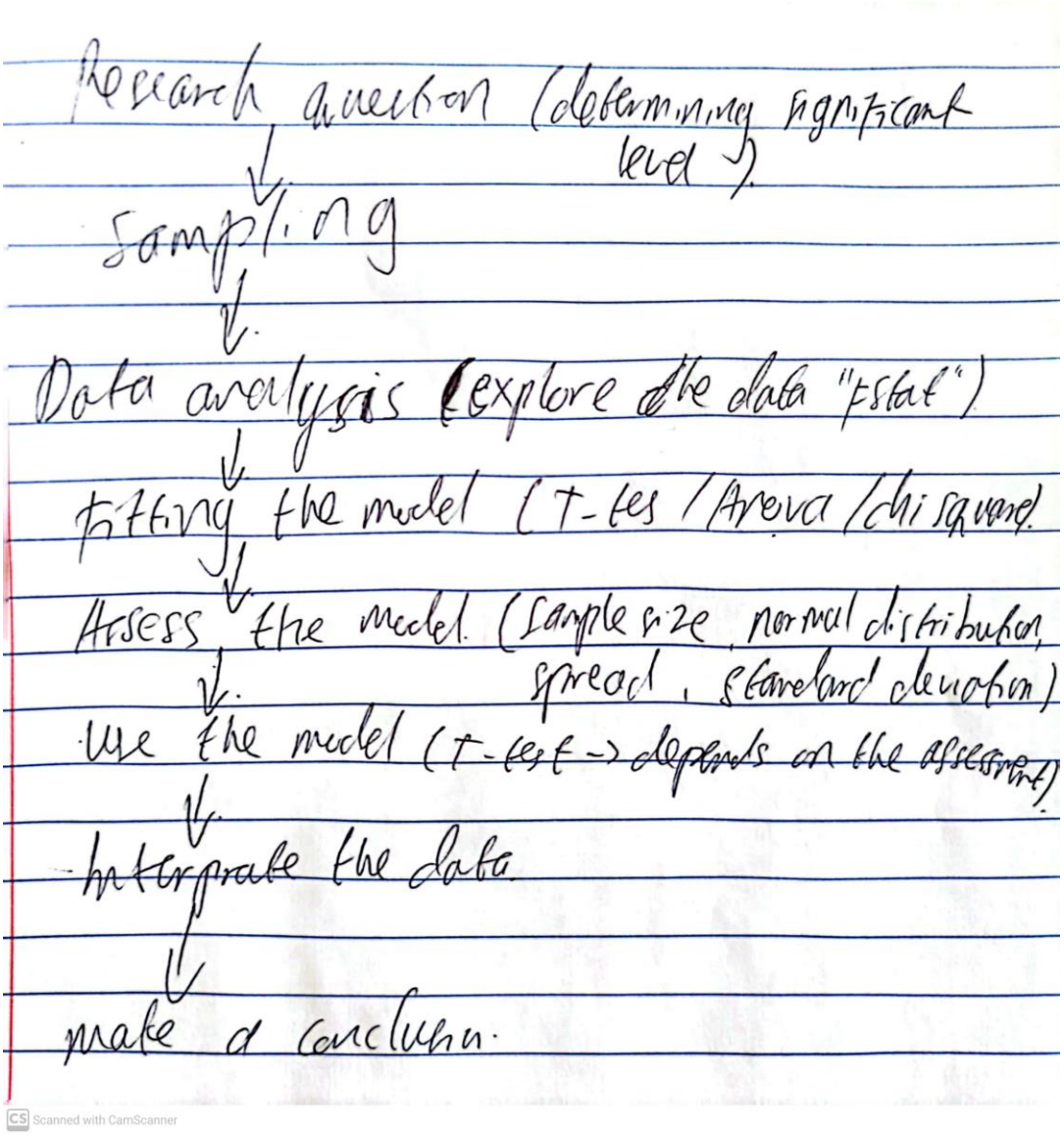


Figure 16

Tal's concept map for the logic of a statistical test

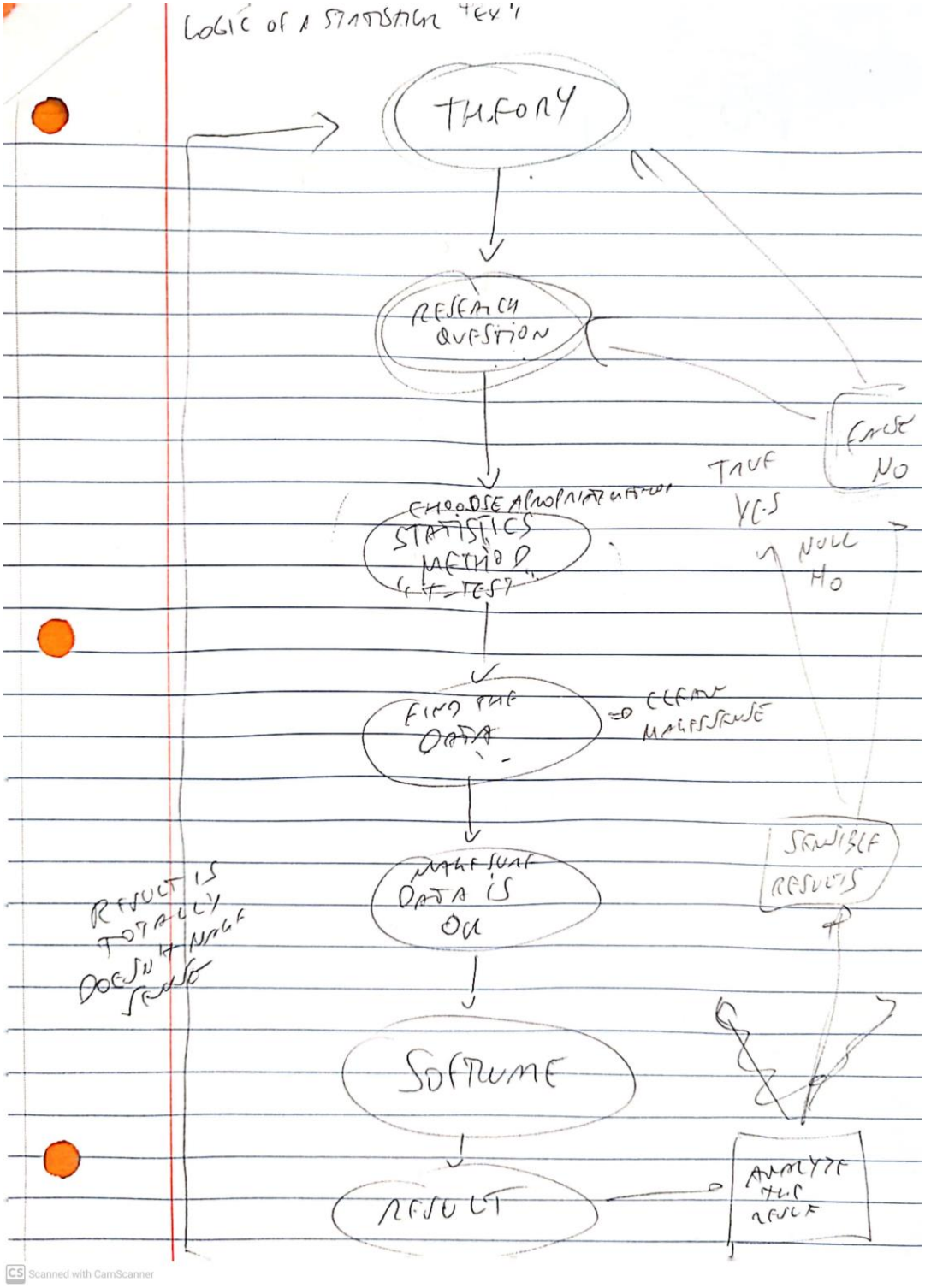


Figure 17

Ade's concept map for the logic of a statistical test

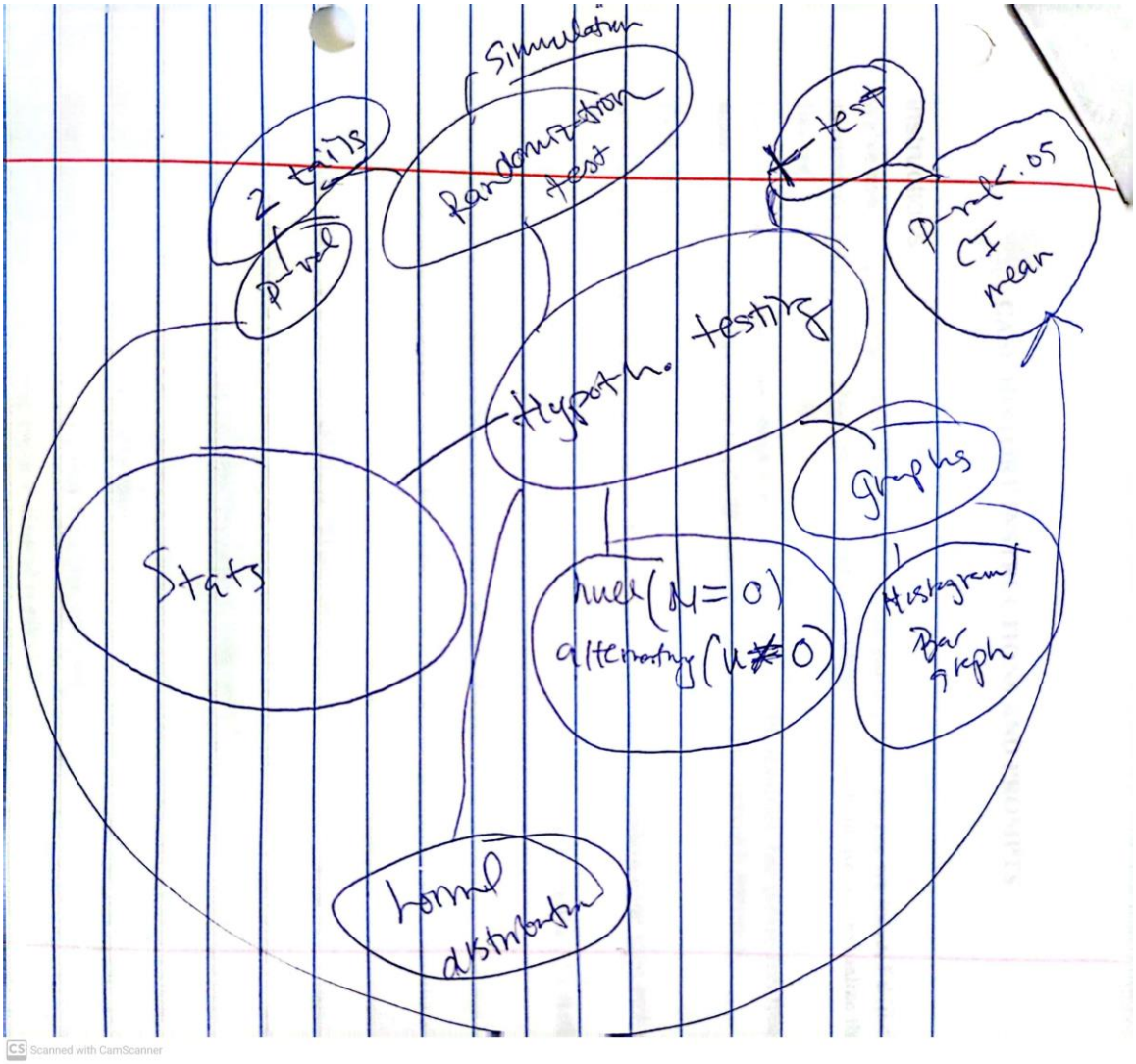


Figure 18

Aan's concept map for the logic of a statistical test

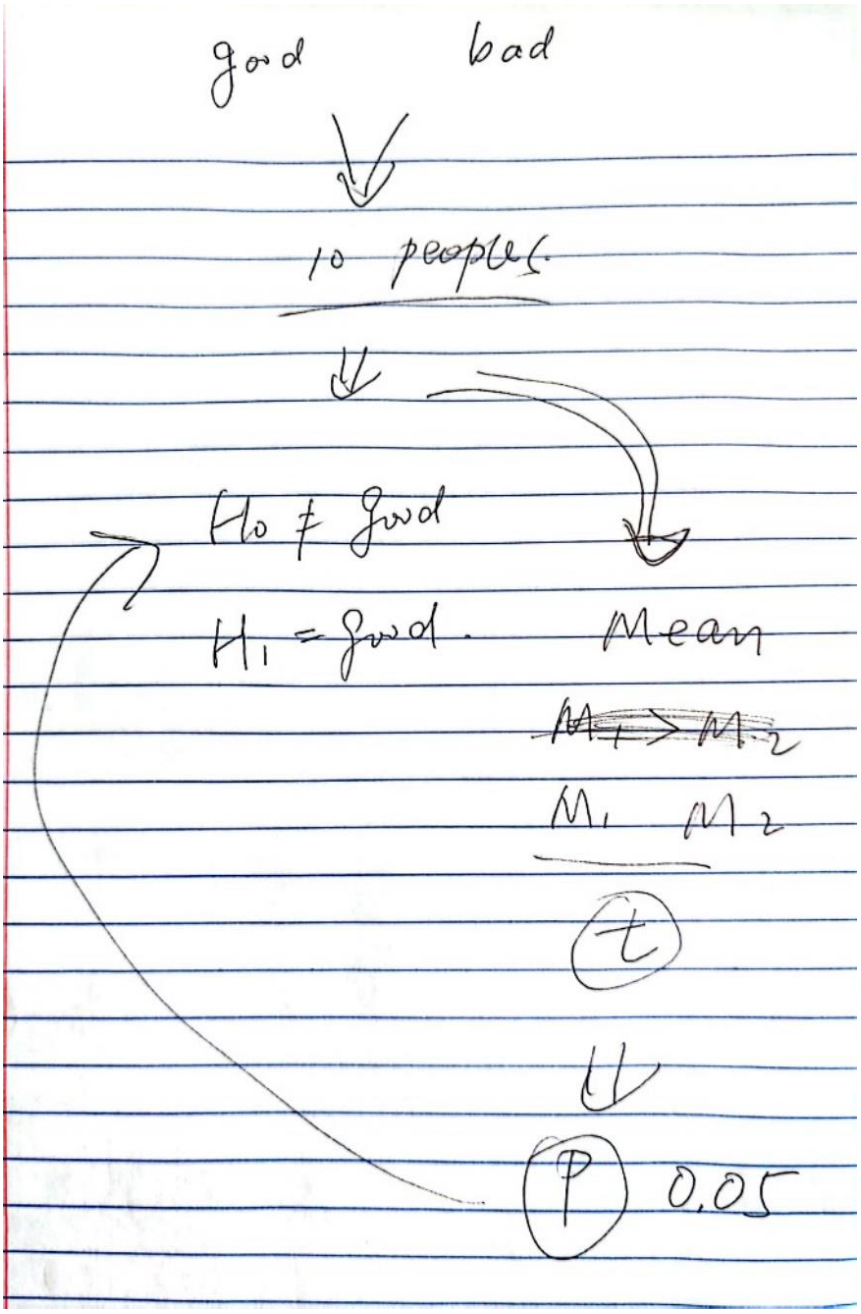


Figure 19

Heat map of the locations on the screen Aan looked at the most while interpreting results from the t-test in R during the Statistical Testing Task (Appendix P01-B1, 13:20 – 14:30), with red indicating a higher amount of gaze for the selected time period , and green indicating lower amounts of gaze for the selected time period

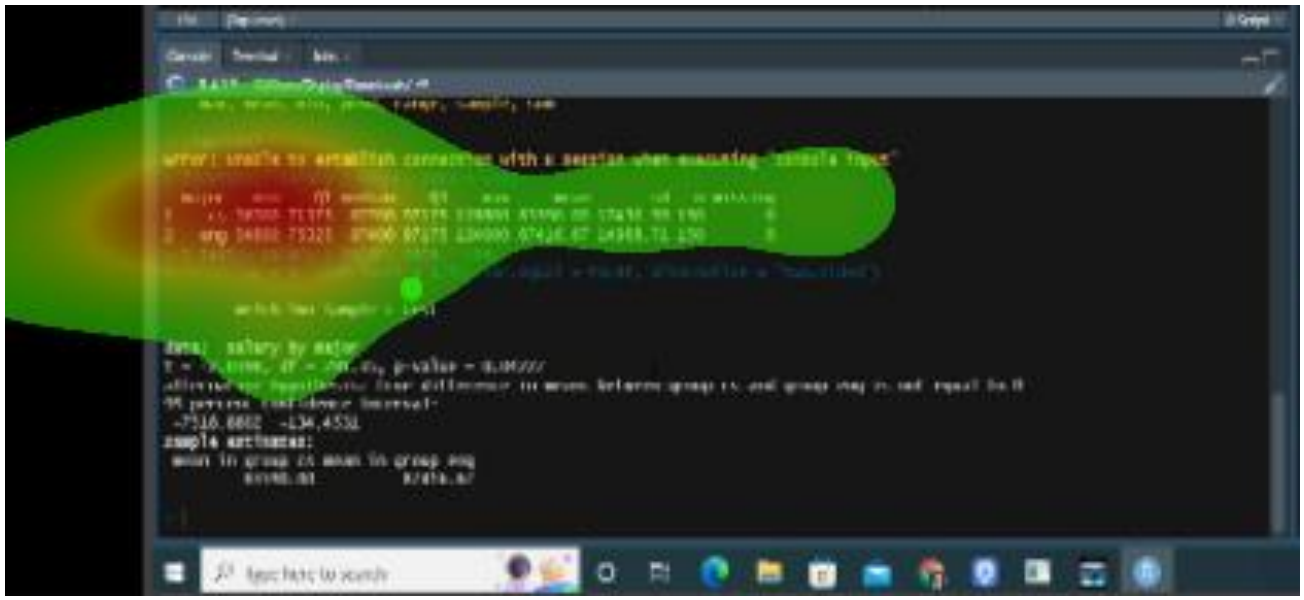


Figure 20

Raw output from R that Aan was looking at while interpreting results from the t-test in R during the Statistical Testing Task (Appendix P01-B1, 13:20 – 14:30).

```
> favstats(data=vse, salary ~ major)
  major   min    Q1 median    Q3    max     mean       sd   n missing
1    cs 36300 71375  82700 97125 128800 83590.00 17430.99 150       0
2   eng 54800 75325  87400 97175 124000 87416.67 14968.71 150       0
> t_test(data=vse, salary ~ major)

        Welch Two Sample t-test

data:  salary by major
t = -2.0398, df = 291.35, p-value = 0.04227
alternative hypothesis: true difference in means between group cs and group eng is not equal to 0
95 percent confidence interval:
 -7518.8802  -134.4531
sample estimates:
 mean in group cs mean in group eng
         83590.00          87416.67
```

Figure 21

Screen shot of the StatKey output that Aan used to think about the Airplane Delays Task as part of the Statistical Testing Task

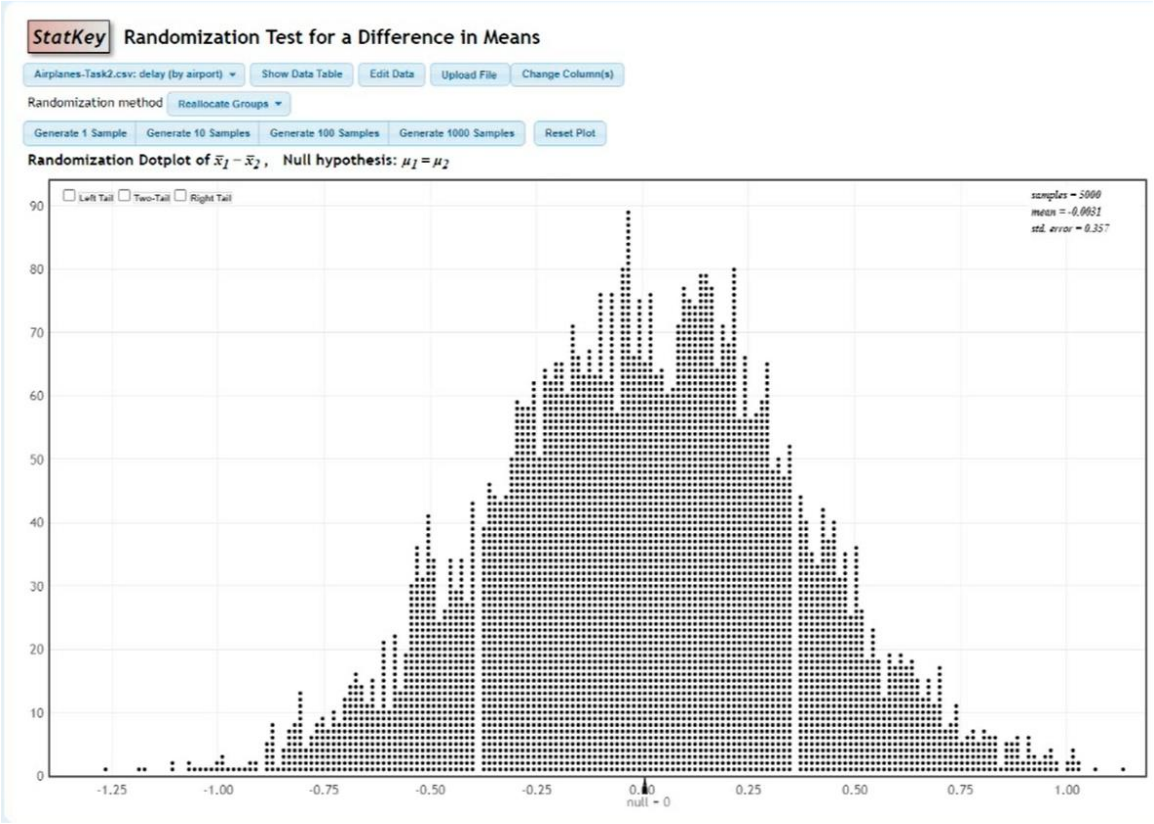


Figure 22

Heat map of the locations on the screen Aan looked at the most while thinking about the question "Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?" during the Statistical Testing Interview, with red indicating a higher amount of gaze for the selected time period, and green indicating lower amounts of gaze for the selected time period

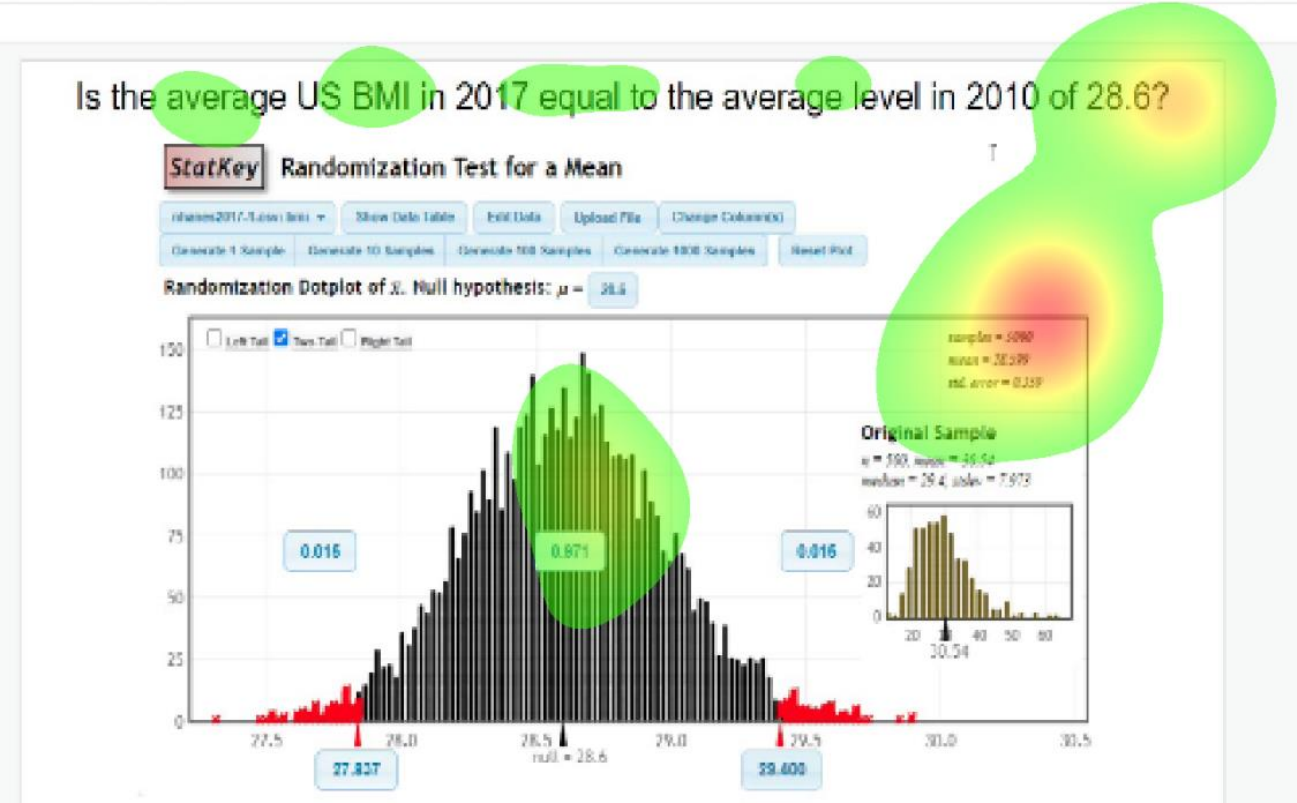


Figure 23

A null model drawn by a member of the EPSY 5261 teaching team, as a response to the question 'What is the one thing that you want students to remember 10 years from now?'

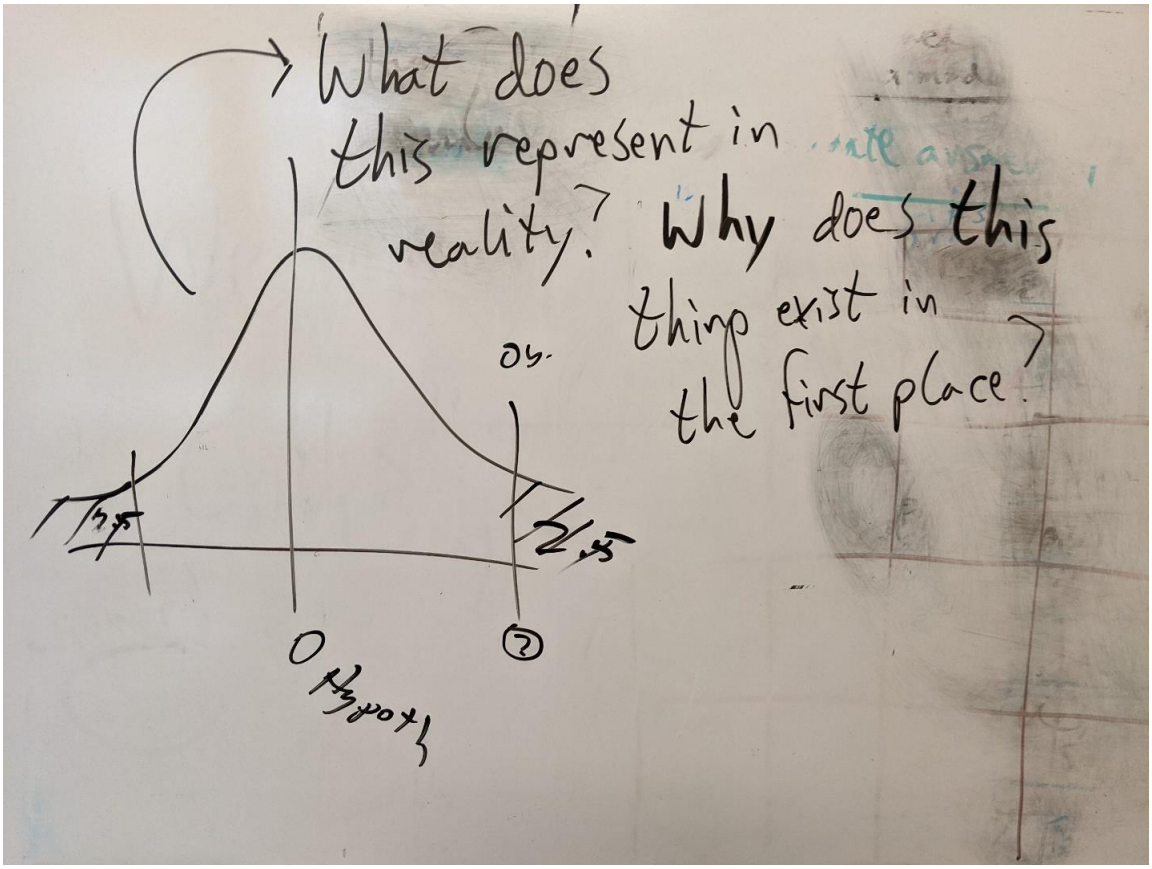
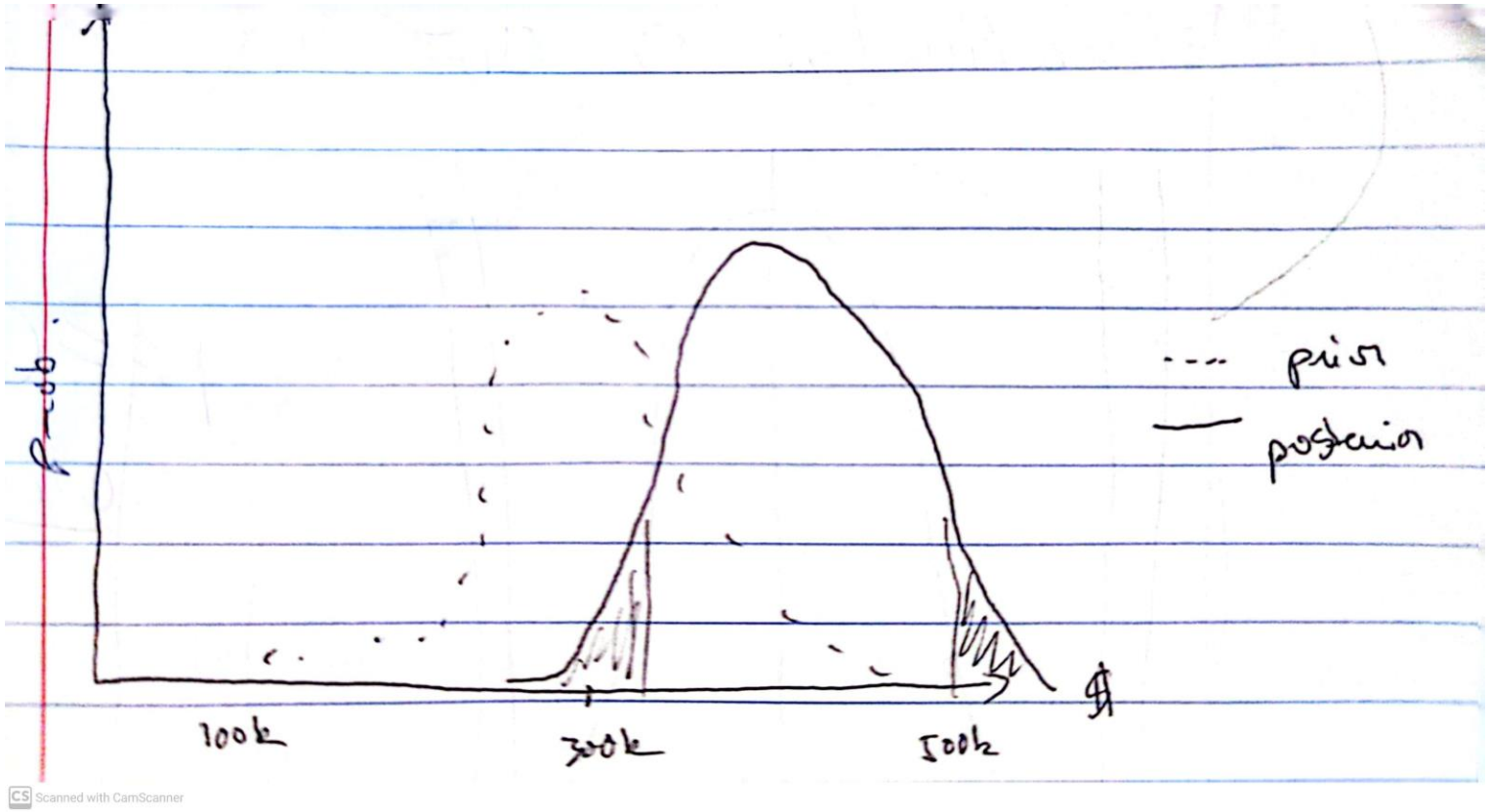


Figure 24

Drawing of a prior and posterior distribution provided to Jaci by the researcher to answer the question 'Is the average home price in New York equal to $300,000?'



Appendix A

Tests of significance items from the Comprehensive Assessment of Outcomes in Statistics (CAOS; delMas et al., 2007)

Item No.	Item Stem	Response Options
19	A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain?	a. A large *p*-value. b. A small *p*-value c. The magnitude of a *p*-value has no impact on statistical significance
23	A researcher in environmental science is conducting a study to investigate the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either a treatment or a control group. The first in the treatment group showed higher levels of the indicator enzyme. Suppose a test of significance was correctly conducted and showed no statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?	a. The researcher must not be interpreting the results correctly; there should be a significant difference. b. The sample size may be too small to detect a statistically significant difference. c. It must be true that the herbicide does not cause higher levels of the enzyme.
	A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with Macular Degeneration. The article gives a *p*-value of 0.04 in the analysis section. Items 25, 26, and 27 present three different interpretations of this *p*-value. Indicate if each interpretation is valid or invalid.	

25	The probability of getting results as extreme as or more extreme than the ones in this study if the drug is actually not effective.	a. Valid b. Invalid.
26	The probability that the drug is not effective.	a. Valid. b. Invalid.
27	The probability that the drug is effective.	a. Valid. b. Invalid.
40	The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternate hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?	a. The circuit is definitely not good and needs to be repaired. b. The electrician decides that the circuit is defective, but it could be good. c. The circuit is definitely good and does not need to be repaired. d. The circuit is most likely good, but it could be defective.

Appendix B

Simulation-based inference topic items from the Goals Outcomes Associates with Learning Statistics (GOALS-4) assessment (Sabbag, 2016; Sabbag et al., 2015)

Item No.	Item Stem and Response Options^	Percent correct*
6	A researcher investigated the impact of a particular herbicide on the enzyme level of carbonyl reductase in fish. In the study, 60 farm-raised fish were randomly assigned to the treatment group (in which they were exposed to the herbicide) or to the control group (in which they were *not* exposed to the herbicide). There were 30 fish assigned to each group. After the study, the data were analyzed, and the results of that analysis are reported in the output below. (p = 0.3644; 95% CI: -11.15 – 4.16) Based on the results of the study, the researchers should not conclude that the herbicide has an effect on the enzyme levels of farm-raised fish. a. Valid b. Invalid	68.3%
14	Two medical researchers each perform the same experiment using two different samples from the same population. One study results in a p-value of 0.06, and the other study results in a p-value of 0.09. Which of the following statements is correct regarding the evidence against the null hypothesis? a. The p-value of 0.06 gives stronger evidence against the null hypothesis because it is smaller. b. The p-value of 0.09 gives stronger evidence against the null hypothesis because it is larger. c. It's impossible to tell which p-value provides stronger evidence against the null hypothesis, because they are both greater than 0.05.	45.2%
	Yolanda was interested in whether offering people financial incentives can improve their performance playing video games. Yolanda designed a study to examine whether video game players are more likely to win a game when they receive a $5 incentive or when they simply receive verbal encouragement. Forty subjects were randomly assigned to one of two groups. The first group was told they would receive $5 if they won the game	

	and the second group received verbal encouragement to "do your best" on the game. Yolanda collected the following data from her study: |	\$5 Incentive | Verbal Encouragement Win | 16 | 8 Lose | 4 | 12 Based on these data, it appears that the \$5 incentive was more successful in improving performance than the verbal encouragement, because the observed difference in the proportion of players who won was (16/20) – (8/20) = 0.40. In order to test whether this observed difference is only due to chance, Yolanda does the following: - She gets 40 index cards. On 24 she writes, "win" and on 16 she writes, "lose". - She then shuffles the cards and randomly places the cards into two stacks of 20 cards each. One stack represents the participants assigned to the \$5 incentive group and the other represents the participants assigned to the verbal encouragement group. - She computes the difference in performance for these two hypothetical groups by subtracting the proportion of winning players in the "verbal encouragement" stack from the proportion of winning players in the "\$5 incentive stack". She records the computed difference on a plot. - Yolanda repeats the previous three steps 100 times.	
15	What is the explanation for the process Yolanda followed? a. This process allows her to determine the percentage of time the \$5 incentive group would outperform the verbal encouragement group if the experiment were repeated many times. b. This process allows her to determine how many times she needs to replicate the experiment for valid results. c. This process allows her to see how different the two groups' performance would be if both types of incentive were equally effective.	34.0%
16	Yolanda simulated data under which of the following assumptions? a. Verbal encouragement is more effective than a \$5 incentive for improving performance. b. The \$5 incentive is more effective than verbal encouragement for improving performance. c. The \$5 incentive and verbal encouragement are equally effective at improving performance	31.1%
17	Below is a plot of the simulated differences in proportion of wins that Yolanda generated from her 100 trials. Based on this plot, the one-sided *p*-value is 0.03.	48.1%

	Which of the following conclusions about the effectiveness of the $5 incentive is valid based on these simulation results? a. The $5 incentive is more effective than verbal encouragement because the *p*-value is less than 0.05. b. The $5 incentive is more effective than verbal encouragement because distribution is centered at 0. c. The $5 incentive is not more effective than verbal encouragement because distribution is centered at 0. d. The $5 incentive is not more effective than verbal encouragement because the *p*-value is less than 0.05.	
18	The *p*-value is the probability that the $5 incentive group would win more often than the verbal encouragement group. a. Valid. b. Invalid.	41.9%
20	In Yolanda's experiment, there were 20 subjects randomly assigned to each group. Imagine a new study where 100 students were randomly assigned to each of the two groups. Assume that the observed difference in this new study was again 0.40 (i.e., that the proportion of wins for the $5 incentive group was 0.40 higher than the observed proportion of wins for the verbal encouragement group). How would the *p*-value for this new study (100 per group) compare to the *p*-value for the original study (20 per group)? a. It would be the same as the original *p*-value. b. It would be smaller than the original *p*-value. c. It would be larger than the original *p*-value.	44.9%

^ Items reported by Sabbag (2016, p. 157-173)

* Percent correct responses from 1,109 undergraduate students from 19 courses in 17 different institutions taken in Fall 2014 (Sabbag et al., 2015)

Appendix C

Recruitment letter sent via e-mail to eligible participants via their instructors

Hello. We are contacting you because you are a graduate student who has completed or is currently enrolled in EPSY 5261. This email describes a study investigating how people think about statistics. The study is being conducted by V.N. Vimal Rao, a Ph.D. candidate in the Department of Educational Psychology at the University of Minnesota, and under the guidance of Dr. Robert delMas and Dr. Andrew Zieffler. Your instructor has given us permission to email you about this opportunity because it is relevant to your training in statistics.

The study asks you to (a) make a concept map, (b) complete two statistical tasks, and (c) a short questionnaire. It takes approximately 90 minutes to complete and pays $50 in the form of an Amazon gift card.

The purpose of this study is to investigate how people complete statistical tests. If you agree to participate, you will meet with the researcher in-person at the Educational Sciences Building to complete the study. You will be asked for your permission to record your gaze with an eye tracking apparatus, and for the meeting to be recorded.

Again, the study will require about 90 minutes to complete and pays $50 in the form of an Amazon gift card.

To participate, you must meet the following requirements:

1. Be a GRADUATE STUDENT at the University of Minnesota
2. Have completed EPSY 5261 in the 2021-2022 academic year

If you do not meet this requirement, then you cannot participate.

If you meet this requirement, and if you are interested in participating, then please reply to this message to rao00013@umn.edu. We will first verify that you meet the requirement above. We will then send you (a) your participant number and (b) a link to schedule a time to complete the study. After completing the study, we will send you the code for an Amazon gift card worth $50.

Thank you for your interest in the study.

Appendix D

Relevant excerpts from an EPSY 5261 course syllabus from the Fall of 2021

Course Syllabus

Introductory Statistical Methods

EPSY 5261-001 – 3 credits
Fall 2021

Audience and prerequisites: This course is intended for upper-level undergraduate and graduate students who have completed a high school algebra course. Although there are no formal prerequisites for this course, students should have familiarity with computers and technology (e.g., internet browsing, Microsoft Word, opening/saving/attaching files, etc.).

Course Description

EPSY 5261 is designed to engage students in statistics by first building a conceptual understanding of statistics through the use of simulation methods and then learning about the more traditional methods, such as *t*-tests, chi-square tests, and regression. This course uses pedagogical principles that are founded in research, such as daily small group activities and discussion.

Attention undergraduates: As this is a graduate level course, it does *not* fulfill the Mathematical Thinking Liberal Education requirement. If you would like to take a statistics course in our department that fulfills that requirement, please consider EPSY 3264.

Course Goals, Objectives and Expectations

Upon completion of this course, students should (1) have an understanding of the foundational concepts of data, variation and inference; (2) be able to think critically about statistics used in popular magazines, newspapers, and journal articles; (3) be able to apply the knowledge gained in the course to analyze simple statistics used in research; and (4) be able to use a statistical software package to analyze data, and appropriately report conclusions from data analyses.

This is *not* a traditional class where you only come each day, listen, watch, and take notes! This class was developed under the inverted classroom model which has a lot of research-based support. The *inverted classroom* "inverts" the traditional instructor-centered classroom model and has you, the student, play a more active role in your learning. You will be required to first read about a topic yourself and complete a short weekly preparation quiz. Then, classroom time will be devoted to learning activities and

discussions to further develop and help you understand the topic. Finally, you will solve problems on homework related to the topic.

This course makes extensive use of *small group activities and large group discussions* to solidify ideas and content, as well as to deepen your understanding of material encountered in the readings. Your learning experience is thus dependent–to some extent–on your classmates and vice versa. Because of this, *it is essential* that you not only attend class each day and participate in the activities and discussions, but that you show up prepared having completed the reading and preparation quiz.

Statistics is more than just an application of mathematics or a methodology used in some other discipline. Statistics is a principled way of thinking about the world. In particular, it is a principled approach to data collection, prediction, and scientific inference.

Statistics is itself a unique discipline that has, like many others, undergone a tremendous amount of growth and change in the last two decades. In today's dynamic and interdisciplinary world, success in confronting new analytical issues requires both substantial knowledge of a scientific or technological area and highly flexible problem-solving strategies.

Internalizing a discipline's way of thinking about and solving problems is a time-consuming process, with the key word being "process". It is not something that can be taught to students in a semester, or even year–long, course. Learning statistics takes much more than memorizing formulae or software commands. It requires active participation and questioning both in and out of the classroom. The instructor of this course will provide you with many opportunities to learn the material through class activities, readings, and homework assignments, but in the end, you will have to do all of the hard work of actually learning that material.

Professionalism: Evidence of professional practice on both our parts includes (a) starting and ending on time, (b) being prepared, (c) being physically and mentally engaged, (d) performing at a high level, (e) making sure cell phones are off, and (f) refraining from sending and receiving e-mail, playing solitaire, shopping, texting, tweeting, and facebooking during class. Thank you.

Textbook and Materials

- ***Statistics: Unlocking the Power of Data*** by Lock, Lock, Lock, Lock, and Lock, **2nd edition EBOOK**:
  - U of M Bookstore https://bookstores.umn.edu/course-materials – search for "EPSY 5261"
    - *Be careful when trying to obtain the book from other sources than above, as some will include WileyPlus access (which is not needed for this course).*
    - *The textbook is not available in print-only format (without the enhanced e-text) from the publisher or the bookstore. If you search online, you may find a cheaper print-only copy, but this may not be much cheaper than the e-text.*

While it is possible to get through this course only with a print copy, you may find it useful to spend a bit more to get the e-text.
    - *The first edition is NOT a good option for you to use in this class. The second edition has some very substantial changes in terms of page/section numbers and content, so the first edition would just leave you very confused and lost! The third edition is very similar to the second edition and is okay to use.*
- **Course Packet (REQUIRED):** The Course Packet will be used on a daily basis in class. This course packet contains the learning activities for the course and can be purchased at the Student Bookstore when you purchase the textbook. You will not need the textbook in class every day, but you will need the course packet for every class session.
- A variety of readings will be provided via the course website throughout the semester. These readings come from different sources–such as journal articles and online resources–and explain terms and concepts, or provide additional information not covered in the textbook. Some of the readings are journal and news articles that report about research studies or data analyses that are related to topics addressed in class activities. These have more detail than you need to know for the course, but they provide real-world examples of the statistical questions and methods you are learning about.

COURSE OUTLINE, TOPICS, AND ASSIGNMENTS

Assignments

Preparation quizzes (25% of your final grade): In a flipped classroom, it is crucial that you come to class prepared, having done the readings for the day. Your preparation grade will consist of your performance on 11 weekly preparation quizzes of 5-6 questions each, worth a total of **25%** of your final grade. These quizzes consist of preparation questions that you will answer based on the readings. The preparation quizzes will be taken on the course website and will be due *before class begins* on the day that they are due. You will be allowed two attempts per preparation quiz, and your grade will be calculated using the higher of the two attempts. Each attempt will last 20 minutes. After the quiz is closed at the beginning of class, you will be able to see your grade, results, and correct responses. Therefore, late quizzes are *not* accepted. Instead, the lowest quiz score will be dropped.

Lab Assignments (30% of your final grade): There will be 4 lab (homework) assignments that together are worth **30%** of your grade. The lab assignments will be completed outside of class (as homework) and submitted electronically via the course website.

As a student of statistics, working through all of the lab assignments is an important piece in building a complete understanding of the concepts, as well as allowing you to practice doing statistics. As a way of connecting the work you are doing across all lab assignments, you will explore the same data set for each lab assignment.

Each lab assignment is set up in Canvas as an online quiz. You can download and view a PDF of the lab assignment. For each lab assignment, you may *choose to work alone or in a group.* Working in a group may allow you to explore answers to a question with other students before submitting your lab assignment. If you work in a group, each individual in the group must submit their own assignment. Your lab assignment should be submitted via the course website before the end of the day that they are due (i.e., by 11:59 PM that day).

***Exams (45% of your final grade)*:** There are two midterm exams and one final exam, which together are worth a total of **45%** of your grade. All of these exams are take-home and are worked through *independently*. In this course, you may use any materials you like to complete exams (e.g., your book, your notes, internet resources, etc.) but you *may not* consult with other people or talk with your peers as you are taking exams. If it is discovered that collaboration has occurred on the exam, you will receive a grade of 0 on that exam.

You will have one week to work on each exam outside of class and then submit your work to the instructor via the course website. The exams will involve using statistical software. More details about the structure of each of these exams will be given in class.

Summary of Assignments

Assignment	Individual or group?	Percent of grade
Preparation quizzes	Individual	25%
Lab (HW) assignments	May work in group, but submit individually	30%
Midterm exam #1	Individual	10%
Midterm exam #2	Individual	15%
Final exam	Individual	20%
Total		100%

CALENDAR

The calendar below lists the tentative dates of the course topics and readings, as well as the tentative due dates for the assignments and exam dates. These dates are subject to change at the instructor's discretion – stay tuned to course announcements. Please note that all Preparation Quizzes are due before class and all other assignments before 11:59 PM on the assigned due date.

Week	Day	Topic (*Book Chapters)*	Activities	Assignments Due
1	1 (September 7)	• Syllabus • Introductions • Introduction to statistical software	• Introduction to Using RStudio for Data Analysis	
	2 (September 9)	• Data collection	• Textbook Scavenger Hunt	

		○ Importance of research questions *(1.1)* ○ Purpose of statistics *(1.2)* ○ Types of studies *(1.3)* ○ Sampling bias *(1.2)*	• Data Collection Articles	
2	3 (September 14)	• Data collection ○ Recall types of studies ○ Scope of conclusions based on type of study *(1.2-1.3)* ○ Random sampling	• Sampling Countries	**Preparation Quiz # 1 (1.1-1.3) due before class**
	4 (September 16)	• Data collection ○ Recall types of studies ○ Scope of conclusions based on type of study *(1.2-1.3)* ○ Random assignment	• Association vs. Causation • Purpose of Random Assignment	
3	5 (September 21)	• Numerical summaries ○ Mean, median, percent, difference in statistics *(2.1, 2.2, 2.4)* ○ Standard deviation *(2.3)* • Technology Reference Guides ○ Entering Data ○ Graphs ○ Descriptive Statistics	• Introduction to Numerical Summaries	**Preparation Quiz #2 (2.1-2.4) due before class**
	6 (September 23)	• Numerical summaries ○ Mean, median, percent, difference in statistics *(2.1, 2.2)* ○ Standard deviation *(2.3)* ○ Resistant statistic *(2.2)* ○ Boxplots and outliers *(2.4)*	• Which Graph has the Larger Standard Deviation • 50 Richest Americans	

4	7 (September 28)	• Introduction to confidence intervals • Confidence intervals using bootstrap techniques (one-sample: *3.1-3.4*) ○ Sampling variability *(3.1)*	• Introduction to Confidence Intervals • Bootstrap Interval M&Ms	**Lab #1 due**

	8 (September 30)	• Confidence intervals using bootstrap techniques (one-sample)
		  ○ Measuring sampling variability: standard error *(3.1)*
		  ○ Constructing bootstrap confidence intervals *(3.3)*
		  ○ Understanding and interpreting confidence intervals *(3.2)*
			Bootstrap Interval: Body Temp	**Preparation Quiz #3 (3.1-3.3) due before class**

5	9 (October 5)	• Confidence intervals using bootstrap techniques: percentile method (*3.4*)
		  ○ Measuring sampling variability: standard error *(3.1)*
		  ○ Constructing bootstrap confidence intervals *(3.3)*
		  ○ Understanding and interpreting confidence intervals *(3.2)*
			• Bootstrap Interval: College Student Debt – Part I	**Exam #1 due by 11:59pm**

	10 (October 7)	• Confidence intervals using bootstrap techniques (paired)
		  ○ Constructing bootstrap confidence intervals *(3.3, 3.4)*
		  ○ Understanding and interpreting confidence intervals *(3.2)*
		  ○ Comparing confidence levels *(3.4)*
			• Bootstrap Interval: Paired Data (Fasting)	**Preparation Quiz #4 (3.1-3.4) due before class**

		○ When to use percentile vs. regular *(3.4)*		
6	11 (October 12)	• Confidence intervals using bootstrap techniques (two-sample, independent) ○ Constructing bootstrap confidence intervals *(3.3, 3.4)* ○ Understanding and interpreting confidence intervals *(3.2)* ○ Comparing confidence levels (using percentile interval) *(3.4)* ○ When to use percentile vs. regular *(3.4)*	• Bootstrap Interval: Comparing Countries (PISA)	
	12 (October 14)	• Introduction to hypothesis tests ○ Purpose of hypothesis test *(4.1)* ○ Null hypothesis and alternative hypothesis *(4.1)*	• Introduction to Hypothesis Testing	**Preparation Quiz #5 (4.1) due before class**
7	13 (October 19)	• Hypothesis tests using randomization techniques (one-sample) *(4.1-4.4)* ○ Intro to p-value	• Randomization test: ESP Study	**Preparation Quiz #6 (4.2-4.3) due before class**
	14 (October 21)	• Hypothesis tests using randomization techniques (one-sample) ○ Conducting randomization tests via applet *(4.4)* ▪ Finding p-values *(4.2)* ▪ Interpreting p-values *(4.2)* ▪ Making conclusions *(4.3)*	• Randomization test: Body Temperature	**Lab #2 due by 11:59pm**

		▪ Significance *(4.3)*		
8	15 (October 26)	• Hypothesis tests using randomization techniques (two-sample) ○ Conducting randomization tests via applet *(4.4)* • Comparing confidence intervals and hypothesis tests *(4.5)*	• Randomization test: Marijuana Users	**Preparation Quiz #7 (4.4-4.5) due before class**
	16 (October 28)	• Hypothesis tests using randomization techniques (two-sample) ○ Conducting randomization tests via applet *(4.4)* ▪ Type I & Type II errors *(4.3)*	• Randomization test: Phone Survey Incentives	
9	17 (November 2)	• Exam 2 Review Day		**Lab #3 due by 11:59pm**
	18 (November 4)	• Normal Distributions *(5.1, 5.2)* • Describing distributions: shape, center, variability • CLT *(5.2)*	• Matching Histograms • Understanding the Central Limit Theorem	
10	19 (November 9)	• Confidence intervals – traditional *(6.1, 6.2)* • One-sample: proportions	• Extra Activity: CI for single proportion (on Canvas website)	**Preparation Quiz #8 (6.1, 6.2) due before class**
	20 (November 11)	• Confidence intervals – traditional *(6.1, 6.2)* • One-sample: means	• Confidence Interval: College Student Debt – Part II	**Exam 2 due by 11:59pm**
11	21 (November 16)	• Confidence intervals – traditional *(6.1-6.4, CI sections)* • Two-sample independent: means	• Confidence Interval: College Student Debt – Part III	
	22 (November 18)	• Hypothesis tests – traditional *(6.1-6.4, HT sections, 6.5)* ○ Two-sample independent: means	• Hypothesis Test: Memory Game	**Preparation Quiz #9 (6.1-6.5) due before class**

		○ Deciding one- vs. two-tailed situations		
12	23 (November 23)	• Chi-square between two variables ○ Test *(7.2)*	• Chi-Square Test: Anemia and Disabilities	**Preparation Quiz #10 (7.2) due before class**
	November 25	**Thanksgiving Weekend – No Class**		
13	24 (November 30)	• Chi-square between two variables ○ Test *(7.2)*	• Chi-Square test:Gallup Poll: US Satisfaction •	**Lab #4 due by 11:59pm**
	25 (December 2)	• Regression: Descriptive ○ Simple linear regression equation *(2.6, 9.1)*	• Regression: Baseball	**Preparation Quiz #11 (2.5-2.7) due before class**
14	26 (December 7)	• Prediction and residuals *(2.6)* Regression: Inference • Data visualization (2.7) ○ Slope & Assumptions *(9.1)*	• Regression: Happy Planet Index Parts I & II	
	27 (December 9)	• Multiple Regression (10.1)	• Happy Planet III • Infant Mortality	
15	28 (December 14)	Final Exam Review Day: Last day of class	• Which Method?	**Review Day**
16	**Final Exam (**December 20**)**	The final is take-home, so you do not need to come to class on this day.		**Take-home Final Exam due by 11:59PM**

Appendix E

Study consent form and information sheet

INFORMATION SHEET FOR RESEARCH
STATISTICAL THINKING STUDY

IRB Code #: 00016330

Version Date: July 14, 2022

You are invited to participate in a research study of how graduate students complete statistical tests. You were selected as a possible participant because you are a graduate student who has recently completed EPSY 5261. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by V.N. Vimal Rao (doctoral candidate, Department of Educational Psychology) under the guidance of Dr. Robert delMas and Dr. Andrew Zieffler.

Compensation:

You will receive payment in the form of an electronic Amazon gift card worth $50 for participating in this study. Your participation is entirely voluntary. You are free to withdraw at any time.

Procedures:

If you agree to be in this study, we would ask you to do the following things:

You will meet with the researcher in person in the Educational Sciences Building. You will make a concept map about statistics, complete two statistical tasks, and a short questionnaire. We will also review the data from the tasks together. You will be asked for your permission to track your gaze with eye tracking apparatus and to record the audio of the meeting.

The session will take approximately 90 minutes to complete.

Confidentiality:

An audio recording will be collected during this study. However, we will immediately use voice filters and modulations in order to mask your voice. We will not store the original voice recording, only this anonymized version. All data analysis will be

conducted on the anonymized audio recording. No other personally identifiable information will be collected during this study.

The records of this study will be kept private. In any sort of report that might be published, no information that will make it possible to identify you as a participant will be included. Research records will be stored securely and only researchers will have access to the records.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota or any other entity or organization. If you decide to participate, you are free to not answer any question or withdraw at any time without affecting those relationships.

Contacts and Questions:

The researcher conducting this study is V.N. Vimal Rao under the guidance of Dr. Robert delMas and Dr. Andrew Zieffler. You may ask any questions you have now. If you have questions later, **you are encouraged** to contact them at:

Mr. V.N. Vimal Rao	Dr. Robert delMas	Dr. Andrew Zieffler
Graduate Student	Professor	Professor
630-999-8118	612-625-2076	612-626-4081
rao00013@umn.edu	delma001@umn.edu	zief0002@umn.edu

This research has been reviewed and approved by an Institutional Review Board (IRB) within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call the Research Participants' Advocate Line at 612-625-1650 or go to z.umn.edu/participants. You are encouraged to contact the HRPP if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

You will be given a copy of this information to keep for your records.

Appendix F

Concept Mapping Task instructions and prompts

Instructions

The first task in this study is about the logic of a statistical test. In this task, I would you to draw/make a concept map for what you think the logic of a statistical test is. There is no right or wrong answers.

I will provide you with pen and paper to make the concept map. If you feel you need to start over, we can begin anew on a different sheet of paper.

As you are drawing your concept map, please explain your thinking outloud. I will also ask you some questions to help get you started, as well as to add in some extra detail to the concept map.

Before we begin, do you have any questions?

Prompts

What is the purpose of a statistical test?

What is the logic of a statistical test?

What role do hypotheses play in statistical tests?

How do you obtain a *p*-value?

How do you compare a hypothesis to evidence?

What is simulation-based inference?

What is the logic of statistical testing in simulation-based statistical tests?

What role do hypotheses play in simulation-based statistical tests?

How do you obtain a *p*-value in simulation-based statistical tests?

How do you compare a hypothesis to evidence in simulation-based statistical tests?

Approaches to statistical testing using T-tests or Chi-Squared tests, for example, are called equation-based inference.

What is the logic of statistical testing in equation-based statistical tests?

What role do hypotheses play in equation-based statistical tests?

How do you obtain a *p*-value in equation-based statistical tests?

How do you compare a hypothesis to evidence in equation-based statistical tests?

How does the logic of simulation-based statistical tests compare to the logic of equation-based statistical tests?

Appendix G1

Statistical Testing Task instructions

In this task, you will be presented with two different problems and presented a research question to answer. You will be told which software application to use in each task – you will use randomizeIt for one task and R for the other. You will also be provided with the necessary data files (in csv format) to answer each research question.

You can answer the question in whichever way you determine is most appropriate. The task is not about whether your answers are right or wrong. Instead, my goal is to get a better idea of how we think when doing statistics.

To help me get an idea of how you are thinking, I'm asking you to *think aloud* as you answer the questions. That means telling me EVERYTHING you are thinking as you read and answer each question.

Please read ALL text and please say ALL of your thoughts out loud. I really want to hear all of your opinions and reactions, negative and positive. Do not hesitate to speak up whenever something seems unclear or is hard to answer.

I'm not here to correct your thinking or guide you; so, if you ask me any questions, I will turn them back to you. I will remind you to think aloud throughout the test. My goal is to keep you talking. I understand that this way of taking a test may feel new or different, so don't worry about whether you're doing well or poorly. That's not what this is about.

Okay, before starting, let's first practice thinking out loud. I will read a question, and I'd like you to think out loud as you answer it. The question is: How do you commute to campus?

[PROBE AS NECESSARY]: Please tell me more about that. Why did you say {answer}?

Before you begin, let me turn on the recording.

Ok, we are now recording. Please proceed when you're ready and begin reading and thinking out loud.

Appendix G2

Instructions for the VSE problem provided to participants as part of the Statistical Testing Task

Difficulty in choosing a career is a common complaint from students. Two careers with approximately the same training requirements are Computer Science and Engineering.

Suppose you have a friend who wants to know if the average salary for Americans who work in the computer sciences is any different from the average salary for Americans who work in engineering.

Using a US Census Bureau database, your friend collected a random sample of 150 salaries for computer scientists in 2019, and 150 salaries for engineers in 2019. Salary is measured in dollars.

Research Question: Is there a difference in the average salary for all Americans between those who are computer scientists and those who are engineers in 2019?

Appendix G3

Instructions for the AD problem provided to participants as part of the Statistical Testing Task

Airplane delays are a common complaint from travelers. Two cities with approximately the same population are Minneapolis-St. Paul and Seattle.

Suppose you have a friend who wants to know if the average delay time for flights from Minneapolis-St. Paul airport (MSP) is any different from the average delay time for flights from Seattle-Tacoma airport (SEA).

Using a US Department of Transportation database, your friend collected a random sample of 150 delay times for MSP airport in 2019, and 150 delay times for SEA airport in 2019. Delay time is measured in minutes.

Research Question: Is there a difference in the average delay time for all flights between those leaving the MSP airport and those leaving the SEA airport in 2019?

Appendix H1

Statistical Testing Interview instructions

In this task, you will be presented with the results of ten different statistical tests. Some results have been generated using simulation-based methods, and some using a *t*-test in R. We'll review each of the tests one at a time.

For each statistical test, you can review the output to get a sense of what is going on. After reviewing the output for a statistical test, please describe the logic or the story of that test. Include an explanation of the hypothesis being tested, the evidence that was collected, and the results of the test.

You can answer the question in whichever way you determine is most appropriate. The task is not about whether your answers are right or wrong. Instead, my goal is to get a better idea of how you think when doing statistics.

Just as in the previous task, and to help me get an idea of how you are thinking, I'm asking you to *think aloud* as you answer the questions. That means telling me EVERYTHING you are thinking as you read and answer each question.

Please read ALL text and please say ALL of your thoughts out loud. I really want to hear all of your opinions and reactions, negative and positive.

Do not hesitate to speak up whenever something seems unclear or is hard to answer. I'm not here to correct your thinking or guide you; so, if you ask me any questions, I will turn them back to you.

I will remind you to think aloud throughout the test. My goal is to keep you talking. I understand that this way of taking a test may feel new or different, so don't worry about whether you're doing well or poorly. That's not what this is about.

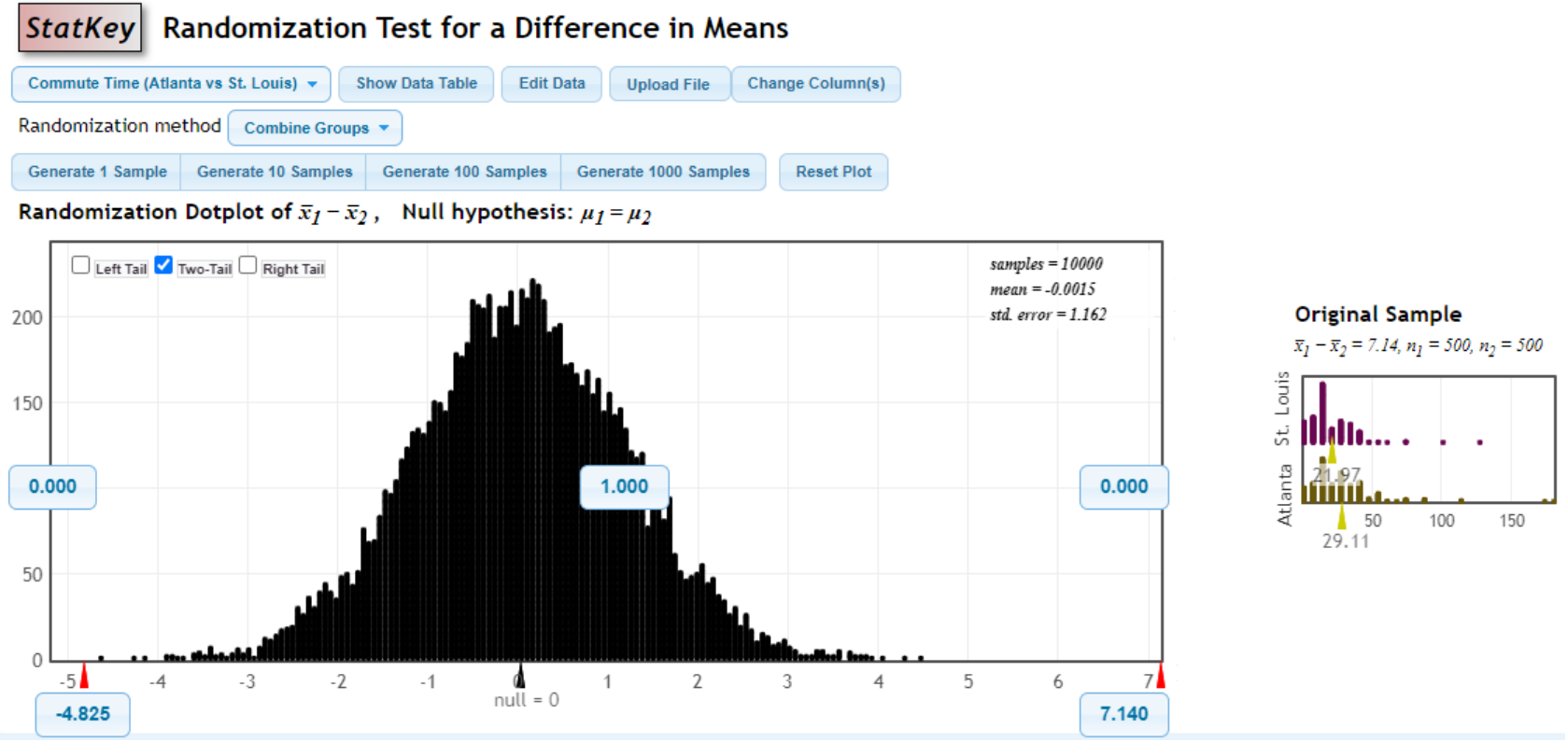
Before you begin, let me turn on the recording.

Ok, we are now recording. Please proceed when you're ready and begin reading and thinking out loud.

Appendix H2

Statistical Testing Interview stimuli

Is average commute time in Atlanta and St Louis the same?



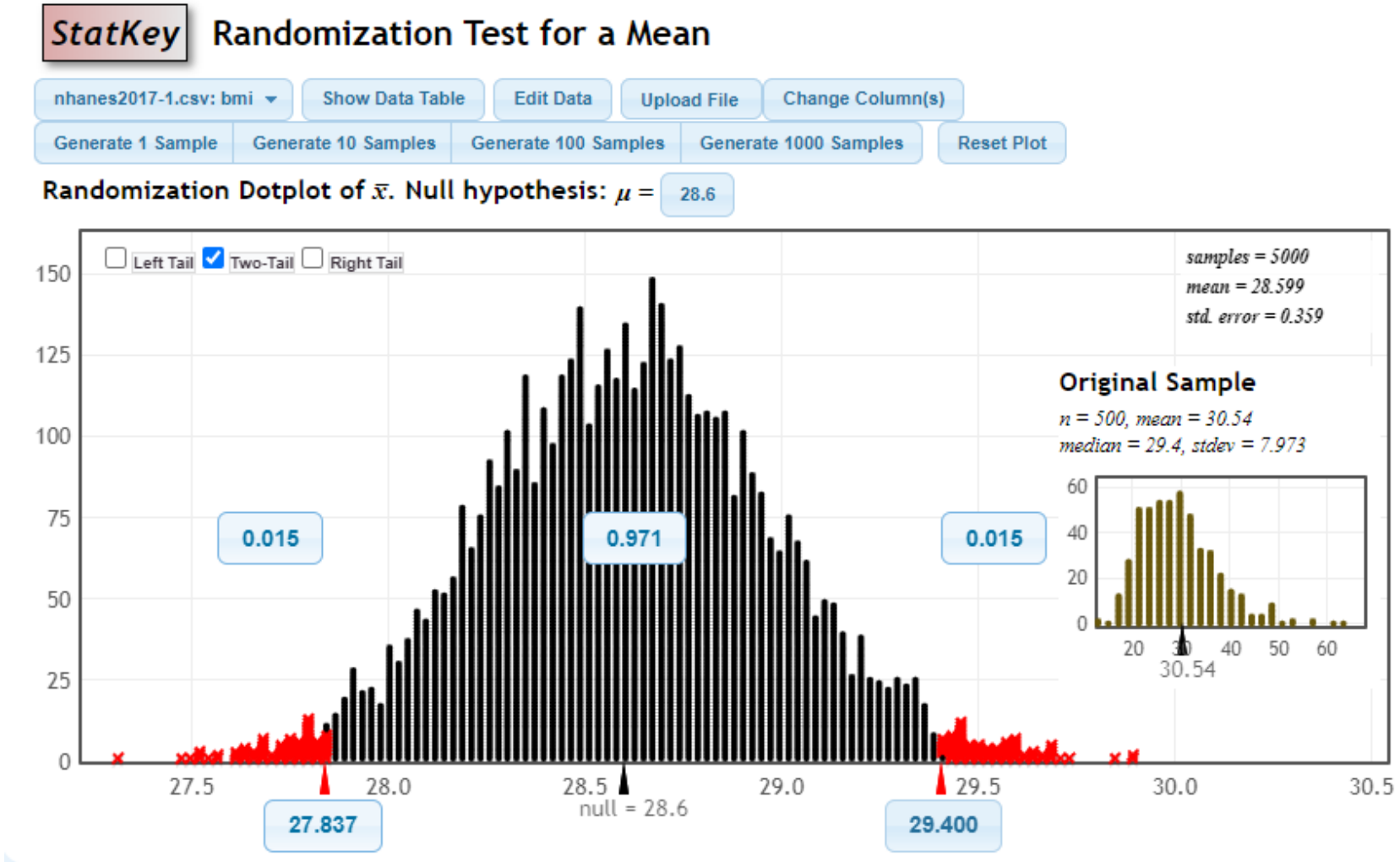
Is the average home price in NYC approximately $300,000?

```
> t_test(data = price, ~ price, mu = 300)

        One Sample t-test

data:  price
t = 2.0856, df = 29, p-value = 0.04592
alternative hypothesis: true mean is not equal to 300
95 percent confidence interval:
 305.1403 826.1264
sample estimates:
mean of x
 565.6333
```

Is the average US BMI in 2017 equal to the average level in 2010 of 28.6?



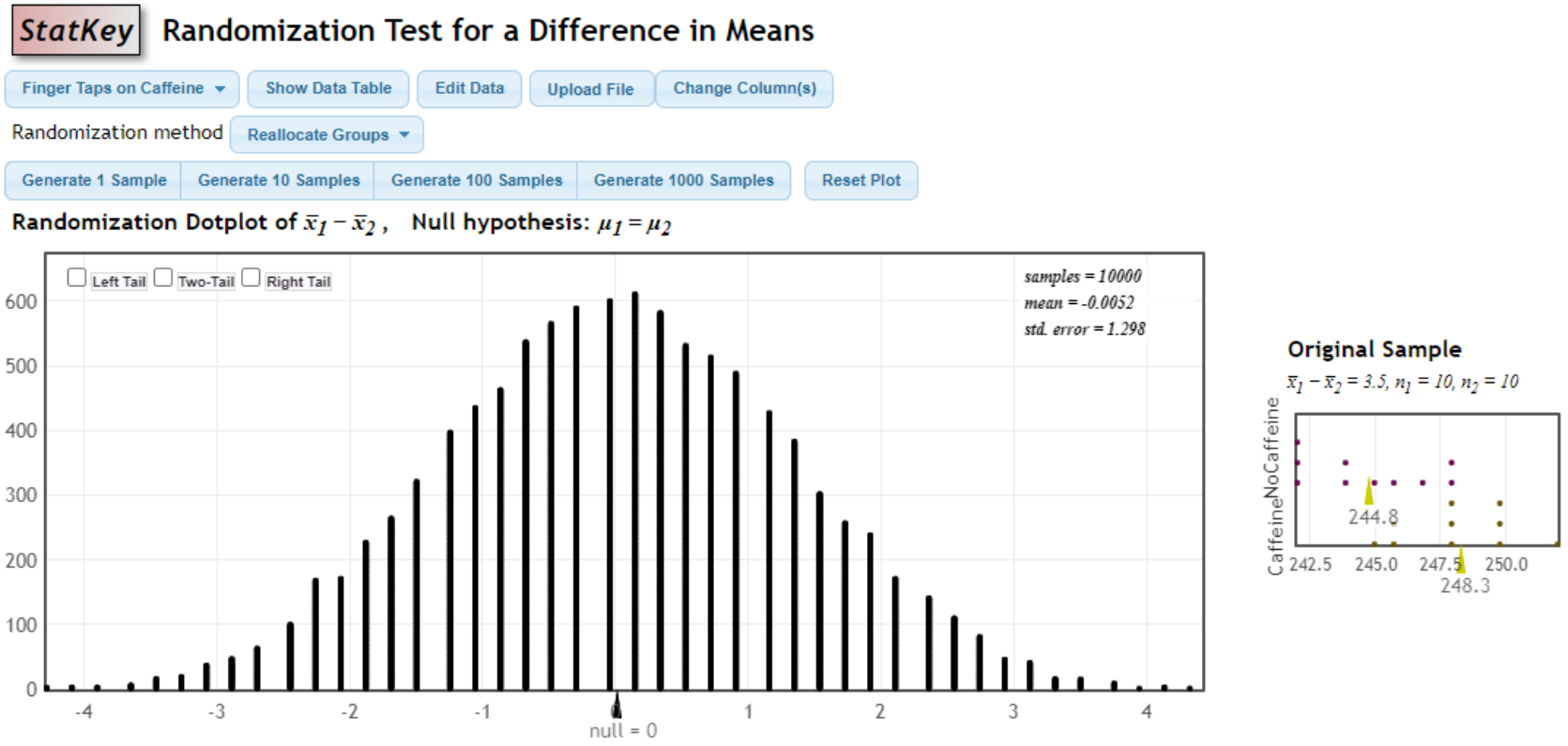
Is the average American family size equal to four people?

```
> t_test(data = nhanes, ~ sizeFamily, mu = 4)

        One Sample t-test

data:  sizeFamily
t = 1.1932, df = 483, p-value = 0.2334
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 2.696553 2.968736
sample estimates:
mean of x
 2.832645
```

Does drinking 2 cups of coffee make you tap your fingers more?



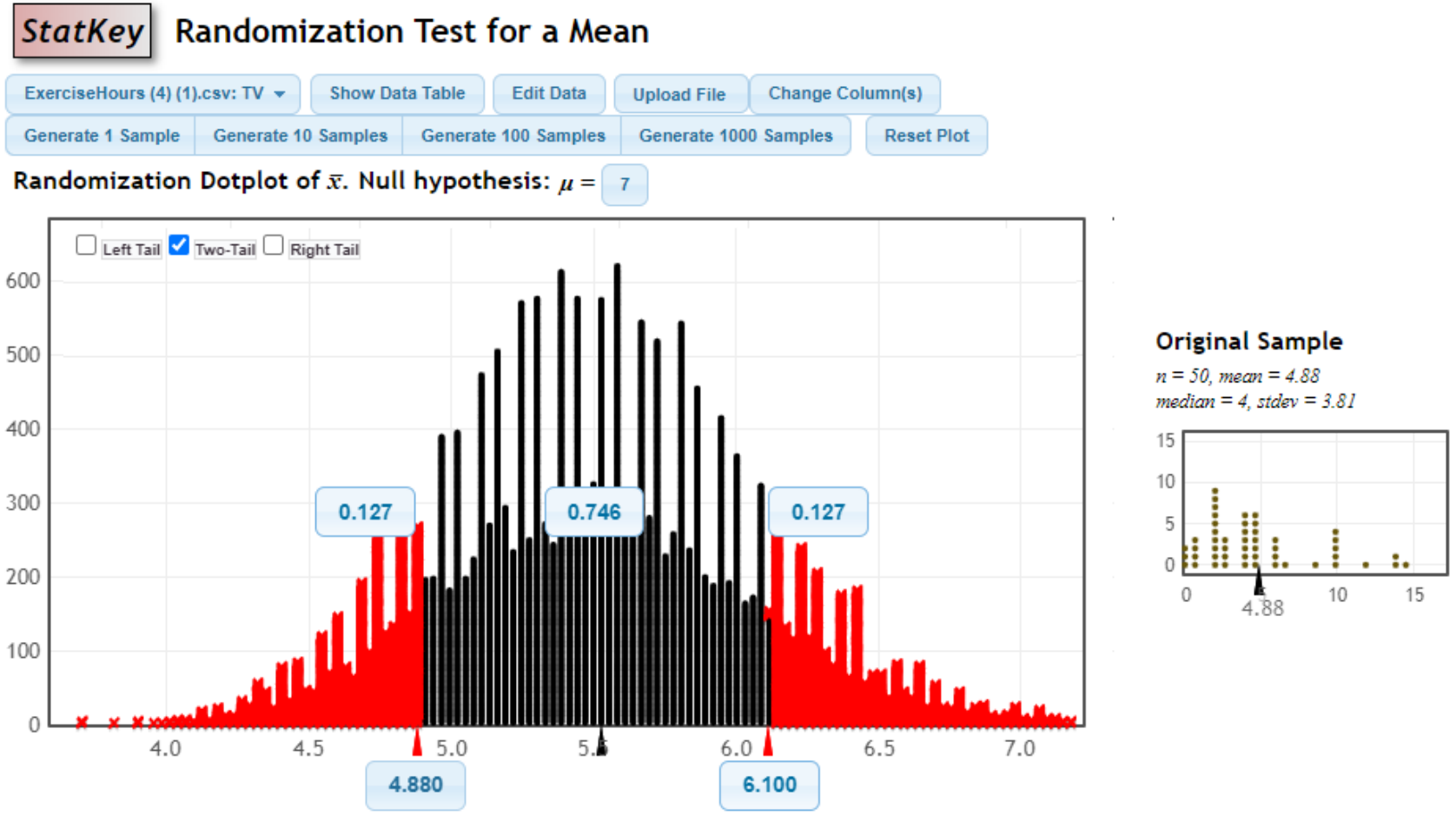
Do UMN students who never feel sleepy exercise more than students who feel sleepy?

```
> t_test(data = ExerciseSleep, Exercise ~ feltSleepy)

        Welch Two Sample t-test

data:  Exercise by feltSleepy
t = 1.2801, df = 19.061
sample estimates:
mean in group Almost Always          mean in group Never
                  105.00000                     72.12121
```

Is the average number of hours that UMN undergrads watch TV per week equal to 7hrs?



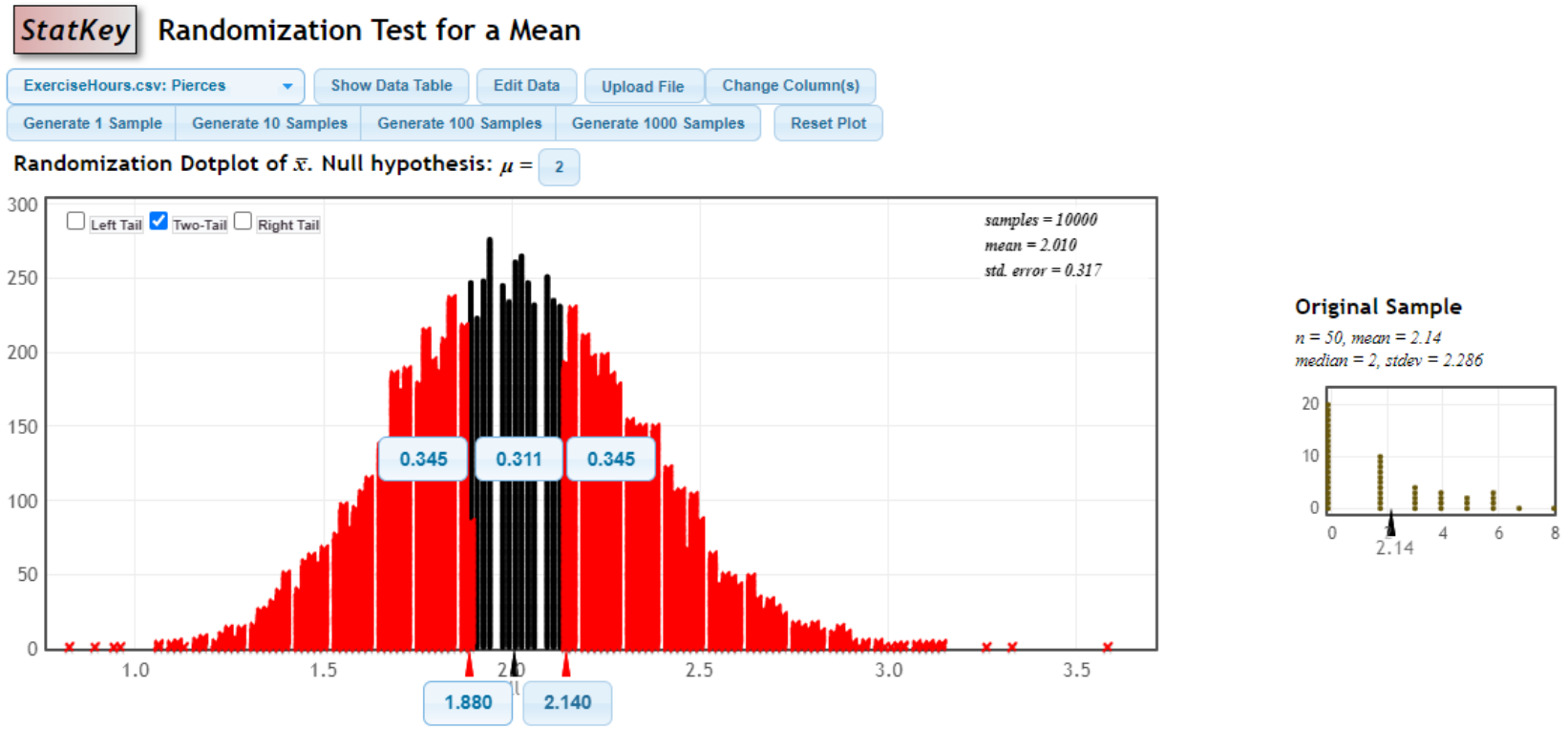
Is Americans' blood pressure related to whether they have trouble sleeping?

```
> t_test(data = nhanes, systolicBP ~ sleepTrouble, mu=0)

        Welch Two Sample t-test

data:  systolicBP by sleepTrouble
t = -0.93179, df = 231.95, p-value = 0.3524
95 percent confidence interval:
 -6.050871  2.165210
sample estimates:
 mean in group NO mean in group YES
         125.6353          127.5781
```

Is the average number of body piercings UMN undergrads have equal to 2?



Is the average PISA Test Score of Finnish students equal to that of Spanish students?

```
> t.test(data=pisa, Total.Score ~ Country)

        Welch Two Sample t-test

data:  Total.Score by Country
t = -0.4225,df = 194.62, p-value = 0.0001544
95 percent confidence interval:
  6.611403 20.428597
sample estimates:
mean in group Finland   mean in group Spain
                81.35                 67.83
```

Appendix I

Video-Cued Interview instructions and prompts

Instructions

In this task, we will watch the recording from the statistical testing task together. The purpose of this task is to add any additional details or comments about your thinking as you completed each task.

You can pause the video at any time to comment on what you are noticing. I will also do this. As we watch the video together, please share as many thoughts as you have about why you did what you did in the video, as well as further details about what you were doing or thinking.

Before we begin, do you have any questions?

Prompts

Can you explain what you were thinking in this clip? What steps did you take, and why?

When were you thinking about hypotheses in this clip?

When were you thinking about *p*-values in this clip?

When were you thinking about the logic of statistical tests in this clip?

When you were looking at [X] in this clip, what were you thinking about?

When you did [A] in this clip, what were you thinking about?