Student Understanding of the Hypothetical Nature of Simulations in Introductory Statistics

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

Jonathan M. Brown

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

Robert delMas, Advisor
Andrew Zieffler, Co-advisor

August 2021

Acknowledgements

Completing the journey of this dissertation and my Ph.D. was the work of numerous people and places! First, thank you to my family for their patience, love, and support. Thank you for not questioning my path even if I communicated great uncertainty on timelines and progress. To my mom, for being the originator of my interest in science and critical thinking, and my first (unofficial) academic advisor. To my dad, for keeping me grounded and never letting me forget about the practical matters of life, and somehow finding a bad joke in every conversation.

Thank you to all my friends. You are the lifelong social vitality that powers me. You have listened to me ramble about things you probably did not care about. More importantly, thank you for not judging me when I slunk away to a corner at house parties to grade and do research while still "being at the party".

I am supremely grateful for my advisors, Dr. Robert delMas and Dr. Andrew Zieffler. You both made me a better writer and thinker and never wavered in providing feedback. Each of you provided support that I know many graduate students do not get and genuinely cared in investing in my progress. There is a difference between getting through a program and growing through a program: your efforts put me on the latter path. Bob, thank you for running with my ideas even when they were well lost in the forest. Andy, thank you for whittling my blocky words and ideas into crisper, more digestible bites.

I want to offer additional deep appreciation to the rest of my committee: Dr. Deborah Levison and Dr. Sashank Varma. Thank you, Deborah, for your feedback and support every step of the way since I graduated from the Humphrey School of Public

Affairs. You helped me along my journey well before I started my doctorate. Sashank, thank you for serving as my committee Chair and offering your unwavering support.

I give special thanks to my statistics education colleagues who offered their expertise and energy when I needed it: Chelsey Legacy and Vimal Rao. You both formed my daily social-intellectual bedrock and provided invaluable contributions to this dissertation study. Chelsey, thank you for helping me pilot the instrument and for the multiple instances of feedback throughout the whole study process. Vimal, thank you for piloting the instrument, your feedback, and carving out those two days back in January to improve my rubric. Above all, thank you, both, for your listening, kindness, and camaraderie across all conversations in the office!

I must also acknowledge all of my statistics education colleagues at the University of Minnesota, including those who graduated before me. I have had the great fortune of knowing and learning from each of you: Michael Huberty, Dr. Elizabeth Fry, Dr. Ethan Brown, Dr. Anelise Sabbag, Dr. Nicola Justice, Dr. Matthew Beckman, Dr. Laura Le, and Dr. Laura Ziegler. (Mike, you and I will forever be cohort-mates!) You all set examples that I have aspired to follow (and put my own spin on). Each of you always made me feel welcome in your presence, professionally and personally.

The statistics education program at the University of Minnesota would not exist without our founder, Dr. Joan Garfield. Thank you for welcoming me into the statistics education family and for creating the structure that has allowed so many students like me to thrive.

I offer additional gratitude to Dr. Isabel Lopez-Hurtado for the countless listening sessions and reflections during writing jams on Zoom. You witnessed me trip through

every little detail in the study design. Your expertise, support, and friendship have made this journey far smoother and happily more memorable.

I give tremendous thanks to the "B-JAMMers" writing group: (soon to be) Dr. Britta Bresina, Dr. Marianne Elmquist, Dr. Andrea Ford, and Dr. Maria Hugh. It is a lot more than a writing group; it is an academic life group. I might well still be swirling in an abstract whirlpool, were it not for your selfless support. Each of you kindly offered valuable slices of your time and energy. Thank you for your ongoing wisdom, humor, friendship, and appreciation of my décor projects!

This work would not have been possible without the instructors and faculty who kindly allowed me to recruit students from their courses: Samuel Ihlenfeldt, Chelsey Legacy, Suzanne Loch, Vimal Rao, Dr. Robert delMas, Dr. Ethan Brown, and Brett Morrow. Thank you for putting energy toward my recruitment efforts amid an already taxing year.

The heart of this study can be found in the humans behind the data. Thank you to all of the participating colleagues and students from EPSY 3264, 5261, and 5262. I am incredibly grateful for your thoughts and words comprising the data of this study. Here's to improving statistics education!

Throughout my entire doctoral path, I have always been able to count on Lori Boucher and Sharon Sawyer, staff for the Department of Educational Psychology. Thank you for being two of the best administrative staff I have ever known. You each kept me on track and guided me to the finish line.

Aside from the people above, numerous other elements have accompanied me on this journey. Each has played a unique and powerful role in helping me cross the finish

line. First, thank you to the Minnesota radio station, KFAN (FM 100.3), for providing the voices in my ear who have kept me company and made me laugh, every day. Second, thank you to the places that provided me with the necessary life balance, which ultimately gave me the energy and focus to navigate the marathon of this degree: The University of Minnesota Recreation and Wellness Center, Vertical Endeavors, Minneapolis Bouldering Project, The Brave New Workshop Comedy Theater, and HUGE Improv Theater. Lastly, thank you to the coffeeshops that kept me afloat and where so much work was produced: UP Café, Starbucks, Caribou Coffee, and Wilde Café.

Dedication

This dissertation is dedicated to my parents, Susanne M. Williams and Robert R. Brown III.

Abstract

Simulations have played an increasing role in introductory statistics courses, as both content and pedagogy. Empirical research and theoretical arguments generally support the benefits of learning statistical inference with simulations, particularly in place of traditional formula-based methods with which introductory students typically struggle. However, the desired learning benefits of simulations have not been consistently observed in all circumstances. Moreover, students in introductory courses have exhibited several types of misconceptions specific to simulations. One theme common to several of these misconceptions is conflating the hypothetical nature of simulations with the real world. These misconceptions, however, have only been discussed in the context of null-hypothesis significance testing (NHST), typically with a randomization-test simulation. Misconceptions about bootstrapping for the purposes of statistical estimation, a common component of simulation-based curricula, have remained unexplored.

The purpose of this study was to explore introductory statistics students' real-world interpretations of hypothetical simulations. The research questions driving this study were the following: (1) *To what extent are there quantitative differences in student understanding of the hypothetical nature of simulations when working with null-hypothesis significance testing vs. estimation?* and (2) *What typical themes emerge that indicate students are conflating the hypothetical nature of simulations with the real world?* The Simulation Understanding in Statistical Inference and Estimation (SUSIE) instrument was created to evaluate student interpretations about the properties of simulations throughout the entire statistical analysis process. The final instrument consisted of eighteen constructed-response items interspersed throughout descriptions of

two different statistical research contexts. One context presented the randomization test for the purpose of NHST, and the second context presented bootstrapping for the purpose of statistical estimation. The instrument was developed, piloted, and updated over eight months and then administered to 193 introductory statistics students from one of two simulation-based curricula. Responses to the instrument were quantitatively scored for accuracy and qualitatively classified for clear examples of conflating the hypothetical nature of simulations with the real world. Quantitative scores were analyzed with descriptive statistics, inferential statistics, and several linear models. Qualitative classifications were analyzed by identifying the primary themes emerging from the responses to each item.

Results from the quantitative analysis suggest that there was no meaningful difference in the aggregate performance between interpreting the randomization simulation vs. the bootstrap simulation (average within-participant instrument section score difference = 0 points, 95% CI: -0.3 to 0.2 points, out of a possible 18 points). However, there was evidence of some differences in performance in parallel items between the NHST and estimation instrument sections. This indicates that participants inconsistently struggled with correctly interpreting the randomization test for NHST vs. bootstrapping for estimation, across the steps of a statistical analysis. Moreover, performance on the instrument overall (average total score = 9.0 points, SD = 3.7 points) and on a per-item basis indicates that several topics were difficult for participants. Bolstering this outcome, results from the qualitative analysis indicate that the participants held a large variety of misconceptions about simulations that pertain to real-world properties and that these misconceptions may vary by the type of simulation used. This

includes thinking that simulations can improve several real-world aspects of studies, can increase the sample size of a study after the data are collected, and are a sufficient replacement for a real-world process such as study replication. Implications from these results suggest the need for real-world conflations to be better addressed in the classroom, a clearer framework to define conflations, and new assessment to efficiently identify the prevalence of conflations and how they emerge when learning statistics with simulations.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The bedrock of the introductory statistics course continues to morph from an emphasis on traditional mathematical formula to technology-focused modern pedagogy and methods. This follows from ongoing calls to update content, pedagogy, and usage of technology in the classroom (e.g., Cobb, 2007; Garfield & Ben-Zvi, 2009; Moore, 1997). Commensurate with these calls, the college Guidelines for Assessment and Instruction in Statistics Education (GAISE) advise instructors to use real data, active learning approaches, and technology for exploring concepts and data (*Guidelines for Assessment and Instruction in Statistics Education College Report 2016*, 2016). Furthermore, arguments have been made to replace or augment the traditional theory-driven statistical curriculum with content and pedagogy based on simulation and modeling (Cobb, 2007; Hesterberg, 1998; Pfannkuch et al., 2018).

In a broad, technical sense a simulation means using a representation of a system or phenomenon for a particular purpose (Banks, 2009; Greca et al., 2014; Ören, 2009). Statistical simulations involve procedures that use representations or models of statistical behavior and processes that employ randomness in some manner (e.g., Freund & Williams, 1966; Mooney, 1997). Simulation methods in teaching statistics may include a buffet of simulation types for several purposes (Chance & Rossman, 2006; Mills, 2002), including learning the central limit theorem (Dambolena, 1986; Mills, 2004), applying and learning statistical resampling methods such as bootstrapping (Engel, 2010; Hesterberg, 1998), and learning statistical inference (Cobb, 2007). Moreover, the scale of change in employing simulations has ranged from new activities (e.g., Chance &

Rossman, 2006) to entirely new textbooks and curricula (e.g., Garfield et al., 2012; Lock et al., 2013; Tintle et al., 2016).

Both theoretical and empirical arguments have been offered for using simulations in statistics courses. Learning theory perspectives include how to manage student cognitive load with simulations (e.g., Budgett & Wild, 2014; Lipson et al., 2006) and how simulations afford learning through constructivism (Erickson, 2006), including guided discovery learning (e.g., Novak, 2014; Reaburn, 2014). Furthermore, some arguments suggest that the characteristics of simulations provide advantages over formula-based approaches to learning statistics, including increasing the transparency of statistical phenomena (e.g., Holcomb et al., 2010), increasing the concreteness of abstract concepts (Chance et al., 2007), and reducing the number of formulae or theoretical distributions down to a few general ideas and procedures (Erickson, 2006; Wood, 2005).

Initial calls for empirical study (e.g., Mills, 2002) have preceded a growing body of literature exploring the learning effects of simulations. Broadly, there is some empirical evidence from quantitative studies that employing simulation-based methods to learn statistical concepts provides learning benefits, particularly above traditional statistical theory-based methods (e.g., Lane & Tang, 2000; Tintle et al., 2014). However, weak or mixed evidence of using simulations instead of traditional approaches to statistics has also been observed (e.g., Beckman et al., 2017; Saputra & Couch, 2018). Importantly, there is no evidence that traditional methods consistently outperform simulation-based methods.

One issue that has emerged from observations about learning inference with simulations is that students appear to develop misconceptions specific to using

simulations (e.g., Case & Jacobbe, 2018; Rossman & Chance, 2014). To organize student misunderstandings of statistical inference with simulations, Case and Jacobbe (2018) proposed a framework illustrating two main areas of student difficulties: (1) problems transitioning between the population, sample, and sampling distribution; and (2) problems coordinating the hypothetical world of the simulation with the real-world of the sample data. Difficulty transitioning between the levels of inference has been previously discussed (e.g., Saldanha & Thompson, 2002). However, the conflation of the hypothetical nature of simulation with the real world remains a largely unexplored area.

Conflating different aspects of simulations with the real world may disrupt statistics learning outcomes and indicate areas where curriculum and instruction should be adjusted. Furthermore, this type of misconception has only been observed or discussed for students answering null-hypothesis significance testing (NHST) or probability questions but not for statistical estimation questions. Conflation-based misconceptions may operate differently when working with NHST vs. estimation statistical research questions. Therefore, additional study of conflation-based misconceptions is needed to characterize their nature, how they vary by statistical task, and how to identify when a student harbors them. Until this work is advanced, the extent of students misinterpreting the hypothetical vs. real-world aspects in a statistical analysis and the consequent effects on learning statistics will remain unknown.

1.1 Description of the study

This study aimed to answer the following research questions:

1. *To what extent are there quantitative differences in student understanding of the hypothetical nature of simulations when working with null-hypothesis significance testing vs. estimation?*
2. *What typical themes emerge that indicate students are conflating the hypothetical nature of simulations with the real world?*

To answer these questions, I developed the Simulation Understanding in Statistical Inference and Estimation (SUSIE) instrument, which was piloted, updated, and administered to one sample of introductory statistics students from courses using one of two simulation-based curricula. The instrument consisted of open-ended items integrated throughout two sections: one section described a statistical analysis that used a randomization-test simulation for a null-hypothesis significance test, and the second section described a different statistical analysis that used a bootstrap simulation for statistical estimation. Responses to the instrument were quantitively scored and qualitatively classified. To answer the first research question descriptive statistics, inferential statistics, and several linear models were estimated using quantitative scores. To answer the second research question clear examples of conflating the hypothetical simulation with the real world were identified using the qualitative classifications.

1.2 Structure of the dissertation

In Chapter 2, the relevant scholarly literature covering several content areas is summarized and multiple frameworks are synthesized. First, simulation and resampling methods are defined generally and as they appear within statistics education research. This includes a synthesis of definitions to propose a framework of the three characteristics that all simulations have. Next, the use of simulations in statistics courses

is considered from the standpoint of several learning theories. This is followed by a review of the quantitative empirical evidence supporting the use of simulation and resampling methods for learning statistics. Next, misconceptions specific to learning statistics with simulations are reviewed. This includes the proposal of a framework based on the synthesis of misconceptions that pertain to conflating the hypothetical nature of simulations with the real world. Finally, relevant gaps in the scholarly literature are summarized to motivate this study.

Chapter 3 describes the methods that were used to execute this study. First, all methods are summarized. Second, the initial development of the SUSIE instrument is presented, including a description of how the contexts, items, and initial blueprint were drafted. Third, the piloting process using Think-aloud interviews and the consequent updates to the instrument are detailed. Fourth, the administration of the instrument in the field test is described. Finally, the data cleaning, quantitative scoring, qualitative classification, and data analysis processes are explained.

Chapter 4 covers the study results in three parts. The first part details the Think-aloud interview feedback and consequent changes that were applied to the instrument throughout the piloting process. The second part describes the results from analyzing the quantitative scores using descriptive statistics, inferential statistics, several linear models, and reliability measures. The third part covers results from the qualitative classification analysis that are relevant to answering the second research question. Finally, Chapter 5 synthesizes the results and discusses the implications. This includes answering both research questions, highlighting key study limitations, and discussing the implications for teaching and research.

Chapter 2: Literature Review

The purpose of this study is to describe and evaluate student understanding of the hypothetical nature of simulations. To frame this study prior scholarly work was reviewed and synthesized. First, the definitions and characteristics of simulations and resampling methods are presented. Second, learning theory arguments for using simulations in statistics education are reviewed. Third, quantitative empirical evidence of using simulations in statistics education is synthesized. Fourth, misconceptions arising from using simulations to learn statistics are summarized. Fifth, implications and extensions of the ideas from the scholarly literature relevant to this thesis are discussed. Finally, the problem statement for this thesis is stated.

2.1 Defining simulation and resampling methods

The purpose of this subsection is to operationalize simulation, in order to frame usage of simulations in statistics education. First, definitions of simulations and a taxonomy of simulation types are presented. Second, the definitions of simulations are synthesized into a set of characteristics. Third, Monte Carlo simulations are defined. Fourth, the nature of resampling methods as a type of simulation is discussed. Finally, the language that describes simulations as they are studied by the field of statistics education research is reviewed.

2.1.1 What is simulation?

"Simulation" is used in many application areas and is defined in myriad ways (e.g., Banks, 2009; Ören, 2011a). At a broad level, definitions of simulations may be split into technical and non-technical meanings. Non-technical meanings, which are observed in various dictionary definitions (see Ören, 2011a), may refer to something imitating

(correctly or incorrectly), being a pretense for, or representing something else (Banks, 2009; Ören, 2011a; Ören, 2009). More concisely, non-technical definitions connect to the adjective, "simulated" (Banks, 2009), as in a simulated pearl imitating a cultivated pearl (Ören, 2009). This review focuses on technical usage. Example definitions from Banks (2009, p. 6) demonstrating the variety of technical simulations include:

- "An unobtrusive scientific method of inquiry involving experiments with a model rather than with the portion of reality that the model represents."
- "A technique for testing, analysis, or training in which real-world systems are used, or where real-world and conceptual systems are reproduced by a model."

To grasp the nature of technical simulations, a funnel approach may be used, where multidisciplinary perspectives provide "wide" or more generalized simulation notions, and examples that are more field-specific provide "narrower" notions. Table 1 provides definitions and descriptions from both a multidisciplinary angle and two narrower angles.

At the widest level, work by Ören (2009, 2011a, 2011b) illustrates that there are numerous definitions and types of simulation and that these both vary by discipline and intended purpose of a simulation. To provide a generalized notion, Ören (2009) offered a simulation definition and a characteristic common to all simulations from a multidisciplinary perspective. Similarly, Sokolowski and Banks (2009) presented a simulation and modeling textbook from a multidisciplinary perspective, including usage external to science and engineering. To this end, Banks (2009) offered a more procedural description.

Table 1 *Simulation definitions from wide and narrow perspectives*

Source	Discipline	Definition
Ören, 2009	Multidisciplinary	Simulation is goal-directed experimentation with dynamic models or use of a representation of a real system to provide experience for entertainment or for training to develop and/or enhance three types of skill, i.e., motor skills, decision-making skills, or operational skills. (p. 155) Use of a representation of reality – whether existing or yet to be engineered, or purely hypothetical – instead of the real system itself is the prominent characteristic of all technical meanings of simulation. (p. 154)
Banks, 2009	Multidisciplinary	To engage [modeling and simulation], students must first create a model approximating an event. The model is then followed by simulation, which allows for the repeated observation of the model. After one or many simulations of the model, a third step takes place and that is *analysis*. Analysis aids in the ability to draw conclusions, verify and validate the research, and make recommendations based on various iterations or simulations of the model. These basic precepts coupled with *visualization*, the ability to represent data as a way to interface with the model, make [modeling and simulation] a problem-based discipline that allows for repeated testing of a hypothesis. (p. 3)
Greca et al., 2014	Science and science education	…[computer] simulations are the representation of the dynamic behavior of a system that moves it from state to state in accordance with an approximate (mathematical) model that is used to implement it on a computer. (p. 900) …all [simulations] may be characterized as transformations of mathematical models in discrete algorithms that imitate the behavior of systems, for which different methods exist to transform the equations into computationally treatable algorithms… (p. 902)

Source	Discipline	Definition
Humphreys, 2004	Computational Science	System S provides a core simulation of an object or a process B just in case S is a concrete computational device that produces, via a temporal process, solutions to a computational model that correctly represents B, either dynamically or statically. If in addition the computation model used by S represents the structure of the real system R, then S provides a core simulation of system R with respect to B. (p. 110)

A definition and a characterization of simulations from Greca et al. (2014) provide a narrower context, given their textual richness and the fact that science and science education are narrower than a multidisciplinary angle, but not as narrow as a single field. Lastly, considering an even narrower, field-specific approach, Humphreys (2004) asserted a criterion-based definition apropos to a philosophy of computational science.

In more field-specific perspectives, the terminology used to define simulation is typically discipline-specific in nature (e.g., “mathematical model”, “computational model”, or “discrete algorithm”). In contrast, multidisciplinary perspectives tend to opt for less restrictive terminology, such as “model”, “representation”, and “observation”. To move beyond field-specific terminology, characteristics of simulations are next operationalized from the wide perspectives.

2.1.2 A taxonomy of simulation types

Work has been done to cut across definitional sensitivity to context and provide a unified framework and synthesis of the characteristics of simulations. Figure 1 presents a visual arrangement of the major components in Ören’s (2009, 2011a) textual summary. There are two major levels of division: technical vs. non-technical forms of a definition,

and experimental vs. experiential purposes of technical simulations. Within an experiential purpose, sub-purposes are divided by training vs. entertainment. If a simulation is for training, then simulation types may be further divided by the intended skill to be trained (motor, decision-making, or operational). Finally, some types of simulations are listed under particular purposes.

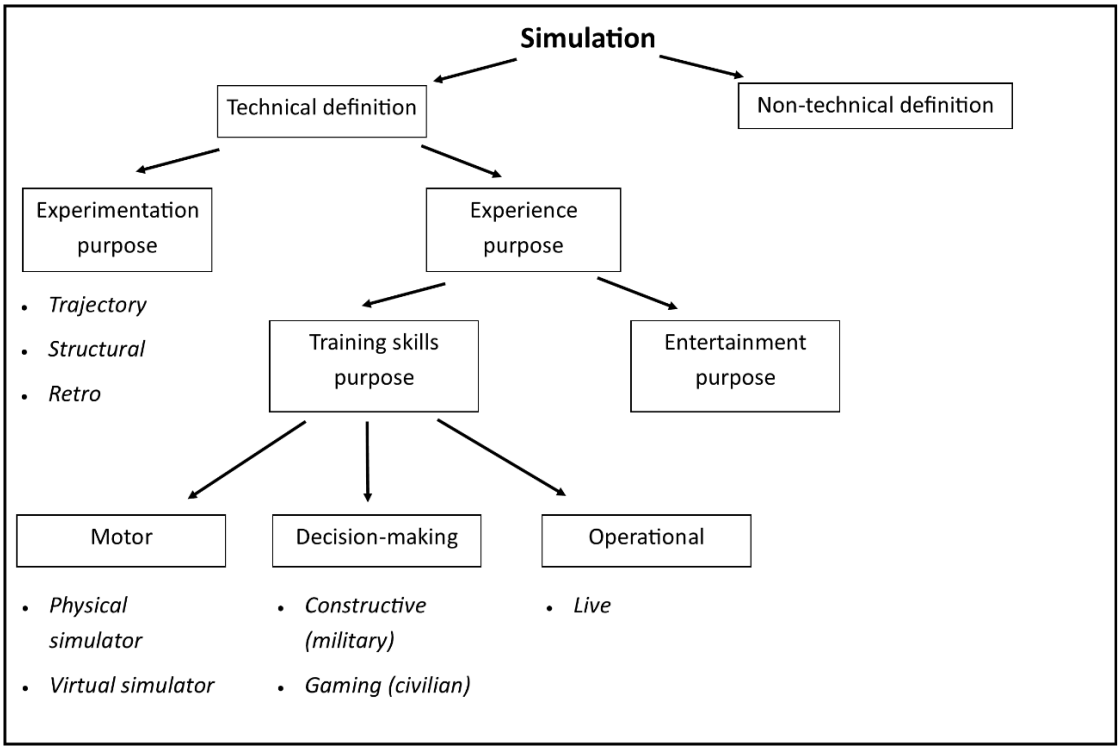


Figure 1. Visual arrangement of select simulation definitions and types from Ören (2009; 2011a). Simulation types are shown as bulleted entries.

Within this framework, the definition of a simulation may be adjusted based on additional factors. Other aspects not shown in Figure 1 but that relate to the arrangement include supplementary features in each branch of the technical side. These include experimentation model types, problem types, and purposes.

Accompanying this framework, Banks (2009) suggested similar categories of division, including a technical vs. non-technical separation. Furthermore, Banks proposed that analysis is an additional broad purpose of simulations. Moreover, Banks suggested a dichotomy of stand-alone vs. integrated simulations, where the former runs independently of the system being represented while the latter runs in conjunction with and to support a real system. For comparison, purposes of stand-alone simulation (Banks, 2009) are similar to those of technical simulation (Ören, 2009), such as education and decision support.

Applying parts of the Figure 1 framework, Ören (2011a, 2011b) categorized nearly 100 simulation definitions across domains and with a consideration of change over time into three groups, each with three sub-categories:

- Modern views of simulations: Experiment, training (for experience), game (for experience)
- Outdated views of simulations: Modeling, model implementation/execution, technique
- Non-technical definitions: Similarity/imitation, pretense/fake, other

The first group reflects the primary division within technical simulations in Figure 1: experiment vs. two types of training, which Ören (2011a) argued as describing a modern view of simulations. The second is exclusively focused on simulations as instances of modeling behavior. Notably, Ören (2009; 2011a) argued against defining simulation as simply modeling, due to model implementation or execution being remnants of outdated methods of creating computer simulations. Similarly, defining a model and engaging in simulation have been described as separate steps in a larger

analysis process (Banks, 2009). In short, not all models or modeling is simulation, but all simulation involves a model. Lastly, the third group captures the latter half of the technical vs. non-technical division from Figure 1. Visual arrangement of select simulation definitions and types from Ören (2009; 2011a). Simulation types are shown as bulleted entries.

2.1.3 Characteristics of simulations

Three characteristics emerging from the three wide descriptions are apparent. First, a purpose is present, in that a simulation is created for a stated purpose. This notion is supported by the fact that numerous purposes exist, which may depend on application area (see Banks, 2009; Ören, 2009). Second, across all three definitions or descriptions, there is an aspect of a model or representation of some entity (e.g., a reality, system, or event). Third, the wide definitions suggest that a simulation is completed by usage, action, or enactment with a model. This additional model behavior varies in terminology and specificity. For example, in the wide definitions, the enactment part of a simulation is non-specifically described as goal-directed experimentation, providing a certain type of experience, or repeated model observation. In contrast, the narrow examples suggest enactment through more exact processes, such as imitating behavior with discrete algorithms or producing solutions to a computational model via a temporal process.

Given that simulations employ an enactment, this implies two additional sub-characteristics to allow the enactment to occur: operationalization and a medium of existence. “Operationalization” speaks to the steps or procedure to execute the simulation. The wide and narrow definitions implicitly refer to procedures in distinct manners. Ören’s (2009) wide definition simply refers to “goal-directed experiments”,

which implies the execution of an experiment, which implies application of a procedure. Banks (2009) places simulation as one step in a larger process or procedure for testing hypotheses of problems. Simulation in this instance is confined to "repeatedly observing a model", and thus, the procedure or operationalization is contained in what observing a model entails. Other parts of the procedure are separated from the simulation operationalization itself, such as analysis and visualization. However, from a narrower perspective, Greca et al. (2014) refer to algorithms that imitate system behavior and that are computationally treatable. In that instance, the algorithm is a set of directions to carry out the simulation.

Medium refers to where the simulation exists. For instance, simulations may or may not involve a computer and thus transcend a specific medium, though their functionality may be dependent on computers. To illustrate: A total of 25 out of 108 reviewed definitions of simulations involved some variation of the words "computer" or "digital", with eight definitions referring to simulation exclusively in a computerized sense (Ören, 2011b). On the other hand, Ören incorporated three definitions where physical or analog aspects were explicitly stated. Concordantly, both Greca et al. (2014) and Humphreys (2004), given their contexts, naturally refer to computer-based simulation. Nonetheless, Greca et al. (2014) recognized that "analogical models" existed prior to the 20th century, and that application with a computer was prominently featured with the introduction of a Monte Carlo method. Furthermore, Geyer (2011, p. 3) argued that "…practical widespread use of simulation had to await the invention of computers", though simulation of random processes had been previously demonstrated. Therefore,

while a simulation needs a medium to exist, the type of medium is a characteristic that depends on the application of interest and the needed tools to execute the simulation.

Table 2 summarizes the characteristics present in the multidisciplinary descriptions and definitions, which can be applied to individual fields or narrower contexts. Thus, how these characteristics are embodied varies by field and application type.

Table 2 *Characteristics of simulations*

Characteristic	Description
1. Purpose	The reasons and desired outcomes for creating and executing a simulation
2. Representation	An entity that represents a real or hypothetical system
3. Enactment	Actions or behaviors with or by a representation, such as a model, that serve as enacting the representation
a. Operationalization	The detailed steps, procedure, or algorithm comprising enactment, to execute a simulation
b. Medium	A medium of enactment, where simulation existence or operation occurs, e.g., computer vs. non-computer

2.1.4 Simulations in statistics via Monte Carlo methods

The application of the characteristics of simulations to statistical simulations invokes the unpacking of Monte Carlo methods. To set the stage, Freund and Williams (1966) define simulation in a statistical context as, “The artificial generation of random processes (usually by means of random numbers and/or computers) to imitate or duplicate actual physical processes.” (p. 101). A notable aspect of this definition is that the representation characteristic of simulations pertains to “physical processes”. Adding

specificity, Freund and Williams provide a cross-reference to a separate entry for Monte Carlo methods.

Monte Carlo methods were first introduced in a physics context to assist with understanding nuclear processes unsolvable via equation, while creating the first hydrogen bomb: "…the method amounted to use of random numbers (*a la* roulette) to simulate the stochastic processes too complex to calculate in full analytic glory." (Galison, 1996, p. 119). Starting in the 1950s as methods were developed, notions of what Monte Carlo methods meant varied by field (for discussion, see Galison, p. 151-152). Nonetheless, Galison argues that the defining quality of Monte Carlo methods is that they are a sampling technique.

The sampling component, as well as other components, are clear in statistical definitions of Monte Carlo methods. Example definitions from a statistical perspective follow:

- "Methods of approximating solutions of problems in mathematics (and related problems in the natural and social sciences) by sampling from simulated random processes…." (Freund & Williams, 1966, p. 66)
- "…using random samples from known populations of simulated data to track a statistics behavior." (Mooney, 1997, p. 2), where "pseudo-population" was additionally stated to describe the entity from which samples are drawn.
- "…any computational algorithm that randomly generates multiple samples of data from a defined population based on an assumed data generating process (DGP). The DGP is the mechanism that characterizes the population from which

simulated samples of data are drawn. Then the researcher explores patterns that emerge across those simulated samples." (Carsey & Harden, 2014, p. 4)

The original notion of Monte Carlo simulations along with their statistical formulations contain specific versions of the characteristics of simulations in Table 2. A general purpose of problem-solving, along with detailed purposes of exploring patterns or tracking statistical behavior are stated. What is represented includes random or stochastic processes related to data samples and populations. Enactment involves employing randomness with a sample and an assumed population, pseudo-population, or data generating process in some manner such as via a computational algorithm.

2.1.5 How is resampling a type of simulation?

In statistics, resampling methods may be regarded as a type of simulation. While resampling methods are applied in a vast array of contexts (see Good, 2006), they are defined exclusively within a statistical context (see Chernick, 2012; Good, 2006). Thus, resampling may be classified as a type of statistical simulation. Accordingly, characteristics of simulations may be applied to resampling methods.

First, resampling may be defined as:

- "[S]tatistical procedures that reuse the sample data for the purpose of statistical inference without requiring parametric assumptions." (Chernick, 2012, p. 255).

Beyond an overall definition, resampling entails a host of methods or procedures. To show this scope, some of these methods with short descriptions of their procedures and purposes are presented in Table 3.

Table 3 *Example resampling procedures and their purposes*

Type	Procedural Description	Example Purposes
Bootstrapping (Chernick, 2012; Efron, 1979; Good, 2006)	Sampling with replacement from data until the sample size is reached, and computing estimates for each iteration of this process	Approximating parameter sampling distribution to assist with estimation, hypothesis testing, and model selection
Jackknifing (Chernick, 2012; Efron, 1979; Good, 2006; Miller, 1974)	Sampling without replacement from data by leaving out a portion of the data (one observation up to and including half the sample), and computing estimates for each iteration of this process	Reducing bias in a serially correlated estimator, estimating parameter variance, hypothesis testing, and model selection
Cross-validation (Chernick, 2012; Good, 2006)	Splitting data into parts, where a model is fit on one part and the fitted model is evaluated on the remaining parts, and computing metrics or estimates at each iteration of this process	Model prediction accuracy and selection, and the degree of smoothing functions
Fisher-Pitman permutation / rearrangement / randomization (Berry et al., 2002; Chernick, 2012; Cobb, 2007; Good, 2006)	Permuting or rearranging sets of grouped data or counts in a contingency table, and computing estimates at each iteration of this process	Hypothesis testing, estimation, and distributional comparison

Based on definitions and descriptions of resampling and the four types in Table 3, characteristics of simulations manifest in resampling methods in particular ways. Table 4 lists each simulation characteristic along with an accompanying explanation of how that characteristic explicitly applies to resampling methods.

Table 4 *Characteristics of simulations applied to resampling*

Characteristic	Application to resampling
1. Purpose	Statistical purposes, such as inference and model selection
2. Representation	Representing or modeling a population or data variability in a population
3. Enactment	
a. Operationalization	Reuse of sample data through one of several methods; may employ a Monte Carlo simulation
b. Medium	May be executed with or without a computer

As indicated in Table 3, resampling methods encompass a host of purposes such as statistical inference or model selection. Irrespective of method, all purposes of resampling are statistical.

Regarding representation, resampling may invoke one or more representations or models of other statistical entities or processes. One form of representation is of a population itself. In analytical statistical inference, a sampling distribution represents theoretically possible values for a parameter, were repeated samples to be drawn from a population. With bootstrapping, for example, instead of actually acquiring a new sample, the original sample operates as a stand-in or hypothetical population from which repeated resampling occurs (Good, 2006; Hesterberg, 2015; Lock et al., 2013). Consequently, the resulting bootstrap distribution of a parameter approximates or represents the parametric sampling distribution (Chernick, 2012).

Another form of representation involves the data variability itself. Again, since the bootstrap procedure invokes resampling from a single sample as a stand-in for the population, the sampling variability of repeatedly drawing from a population is mimicked (Lock et al., 2013). Similar representation is observed with permutation or randomization

tests, where grouped data are repeatedly reordered, such as in a randomized experiment. As explained by Cobb (2007), “The model simply specifies that the observed values were randomly divided into two groups. Thus there is a very close match between the model and what actually happened to produce the data.” (p. 4). Cobb (2007) additionally calls the two groups in this scenario the “simulated” control and treatment groups, further emphasizing their role as representations of the actual control and treatment groups. Rerandomization in this instance produces a set of simulated alternate realities, mimicking and capitalizing on the chance variation used in producing the original data.

To further clarify the representations involved with resampling, considering the enactment characteristic of simulations via Monte Carlo method is helpful. Regarding enactment, all resampling methods are operationalized by reusing sample data. However, since each resampling method consists of a specific sequence of steps, the exact operationalization varies by how the sample data are reused. Revisiting the statistical simulations discussed prior, one notable form of resampling operationalization is with Monte Carlo simulation.

For some resampling methods, the theoretical result of the method is a distribution of a statistic from all possible resampled units arising from a given sample (see examples of this for the permutation test in Chernick (2012) or Good (2006)). That is, once a sample is chosen, it is a matter of computational or manual effort to find all possible rearrangements or resampled units, leading to the exact distribution of interest. Despite this, for sufficiently large samples enumerating all possible resampled units becomes computationally intensive or impractical, and an approximation of the exact resampling distribution may be created from a subset of all possible resampled units

(Ricketts & Berry, 1994). Using the bootstrap method as an example, a given sample has a predetermined set of bootstrap samples and thus a predetermined bootstrap distribution. In addition to systematically listing each possible bootstrap sample, Efron (1979) showed that such a distribution could be computed via formula. On the other hand, a subset of bootstrap samples may be found by random resampling with replacement for a specified number of iterations, i.e., via a Monte Carlo approximation or simulation (Chernick, 2012; Efron, 1979; Engel, 2010; Good, 2006). The collection of computed values from each iteration then forms an "empirical" distribution (Chernick, 2012; Efron, 1979; Good, 2006) that approximates the exact distribution, which itself is an approximation of the sampling distribution.

Since resampling is a type of simulation that may be operationalized via another type of statistical simulation, ambiguity of characteristics may result from this apparent nesting. Monte Carlo simulations involve representations and enactments pertaining to randomness, samples, and populations or data generating processes. In contrast, resampling methods are characterized with respect to samples as proxies for populations. As Carsey and Harden (2014) discuss: "Monte Carlo simulation and resampling methods have a lot to do with the phrase 'in repeated samples'." (p. 2). To distinguish: "Resampling techniques are similar in that they also draw multiple simulated samples of data. However, rather than making draws from a theoretical population [Data Generating Process] defined by the researcher, resampling techniques draw multiple simulated samples from the researcher's actual sample of data." (p. 4).

Thus, both Monte Carlo simulations and resampling methods are executed with data of some kind, but the former is considered to use real or artificial data that represent

a population whereas the latter operates with sample data. Nonetheless, this difference in data types is reconciled by assuming that when a Monte Carlo simulation is applied in a resampling process, the sample represents some population.

Despite potential differences in how the data may be defined, the usage of Monte Carlo simulation in resampling actually emphasizes the representational aspect of resampling. Mooney (1997) referred to Monte Carlo simulations as drawing from a "pseudo-population". This reflects how resampling treats a sample as a stand-in population. Additionally, since Monte Carlo simulations are described as sampling from a simulated random process (Freund & Williams, 1966) or a population based on a data generating process (Carsey & Harden, 2014), this supports the notion that resampling methods are representing or modeling sampling and data variability that would have been observed in reality. The nature of Monte Carlo simulation takes advantage of representations and models of acquiring data from populations, which can then be used in a resampling method.

Lastly, the medium in which resampling methods are enacted may either be by computer or by hand. Exemplifying this is the fact that the permutation method was developed prior to computers (e.g., Pitman, 1937), and thus, exact distributions could not be produced with computers. Several examples of carrying out a permutation or randomization procedure without a computer are available (see Cobb, 2007; Good, 2006; Lock et al., 2013).

2.1.6 What terms refer to simulations in statistics education research?

When simulations are used for educational purposes in a broader multidisciplinary sense, they may be classified as "stand-alone" (Banks, 2009) or "experimental"

simulations (Ören, 2009). As simulations are studied in statistics education research, terminology varies.

Employing the term "simulation" in statistics education may invoke different methods or tests, depending on the context and authors. In broadest usages, "computer simulation methods" (Mills, 2002) and "simulation approaches" (Wood, 2005) have been employed as catch-all terms for any of several statistical methods or ideas, including resampling methods. In other cases, terminology is differentiated, such as simulation versus bootstrapping (e.g., Hesterberg, 1998), or simulation versus resampling (e.g., Biehler, 1997). In the latter case, Biehler (1997) grouped simulation and resampling under a broader "Monte Carlo workbench" for using such methods. Similarly, Simon et al. (1976) implemented the "Monte Carlo method" as a catch-all term for an unidentified set of statistical problems, but this was prior to Efron's (1979) influential work on bootstrapping. In at least one other instance, the terms "simulation", "randomization", and "resampling" are employed as separate terms, but are not explicitly differentiated by the authors (see Stephens et al., 2014).

The usage of "randomization" as a term connected to simulations has varied. Randomization may generically refer to set a of methods involved with the act of randomizing data (e.g., see Fitch & Regan, 2014; Lock et al., 2013), the phenomenon underpinning both random sampling and assignment (see Cobb, 2007), or a single inferential test or method (e.g., see Chance & Rossman, 2006; Pfannkuch & Budgett, 2014). Moreover, the randomization test as a single entity has been subject to alternative terminology, as it is also known as the permutation test (Budgett et al., 2013; Erickson, 2006; Taylor & Doehler, 2015).

What constitutes simulation is further clouded when considering curricula. Statistics curricula naturally vary by course and institution. Thus, different simulation-based courses may include non-overlapping sets of course components and dissimilar descriptions. For instance, "simulation-based curricula" have been implemented with curricular components of bootstrapping for confidence intervals and simulation-based ("randomization") tests for inference, with a variety of software (Chance et al., 2016; Hildreth et al., 2018; Maurer & Lock, 2016). On the other hand, a course with similar content and pedagogy was initially called a "randomization-based" curriculum (Tintle et al., 2012; Tintle et al., 2011). Yet another variation: The Change Agents for the Teaching and Learning of Statistics (CATALST) curriculum "…uses the ideas of chance and models, along with simulation and randomization-based methods…" (Garfield et al., 2012), implemented with the *TinkerPlots*^{TM} software (Konold & Miller, 2011). Therefore, a course title or curriculum title focused on simulations may not signal consistent usage or inclusion of the same software, topics, or methods.

2.1.7 Summary of how to define simulation in statistics education

Based on definitions and conceptualizations from multiple fields, every simulation may be described from the basis of three characteristics: purpose, representation, and enactment. In statistical contexts, these characteristics take on specific meanings. Purposes are inherently statistical, such as for conducting inference. Representation includes the modeling of statistical phenomena, such as sampling variability. Enactment invokes a range of procedures that consist of different steps (i.e., operationalization) through computers or manual by-hand operations (i.e., mediums). A common class of statistical simulations are those from resampling methods, which all

operate with the same approach to enactment and representation: working with sample data in some fashion as a pseudo-population or proxy for a larger population.

Simulations as they are used in educational settings fluctuate in how they are identified and used. While a typical set of methods appear in the literature, such as randomization and bootstrapping, there is nonetheless variation in the exact types of simulations employed in different educational settings and some names for simulations mean different methods or sets of methods in different studies. Such variability is especially apparent when comparing simulation-based curricula.

2.2 The role of learning theory

The learning benefits of simulations have been suggested based on a host of principles and intuitive arguments (see e.g., Chance et al., 2007; Cobb, 2007; Pfannkuch & Budgett, 2014), and by drawing from theories of learning, with some concern for learning drawbacks. Such arguments as to why simulations are beneficial for student learning primarily stand in contrast to non-simulation-based methods. Key learning theories in this regard include constructivism (e.g., Ben-Zvi, 2000; Erickson, 2006; Wood, 2005), discovery learning (e.g., Lane & Peres, 2006; Lipson et al., 2006), cognitive load theory (e.g., Budgett & Wild, 2014; Chance & Rossman, 2006), and schemas (e.g., Lipson, 2002; Lipson et al., 2003), among others.

The reasons why learning benefits may be expected from a learning theory standpoint are next detailed. First, key learning theories are introduced. Second, simulation will be argued to be a form of guided discovery learning. Third, examples of where these theories have been tied to simulations in statistics education are detailed.

Lastly, linkages between learning theories and the characteristics of simulations are discussed.

2.2.1 Overview of key learning theories

Schemas, hypothesized to be foundational for how people learn, are patterned, organized structures of information that move thinking from working memory to long-term memory (Centre for Education Statistics and Evaluation (CESE), 2017; National Research Council (NRC), 2001). Along these lines, learning may be defined as constructing numerous schemas in long-term memory (Sweller, 2011) and then employing such schemas in an automatic way (de Jong, 2010).

To build schemas and encode them into long-term memory, the cognitive load must be managed. Cognitive load may be defined as that which results from processing new information in working memory (CESE, 2017), a construct for how the combination of learner and task characteristics impose a load on cognitive structures (Sweller et al., 1998), or the capacity of working memory needed for a learning task (de Jong, 2010). To maximize learning, pedagogy should seek to minimize extrinsic load, while being mindful of intrinsic load and maximizing germane load (for a review of these load types see CESE, 2017; de Jong, 2010; Sweller et al., 1998). At a practical level, cognitive load is influenced by the "element interactivity", or the degree to which multiple elements must be simultaneously managed in order to learn a concept (Sweller, 2011; Sweller et al., 1998). Framing learning from a cognitive load perspective suggests an exclusive focus on direct, explicit, or guided instruction over minimally-guided or constructivist forms of instruction (e.g., CESE, 2017; Kirschner et al., 2006).

Constructivism advocates for the position that learning occurs when learners actively construct knowledge (de Jong & van Joolingen, 1998; Mayer, 2004; Schwartz & Bransford, 1998), and that all learning requires a constructivist process of some kind (Hmelo-Silver et al., 2007). A common form of constructivism is discovery learning, which may broadly be defined as discovering knowledge for oneself (Bruner, 1961) or with limited guidance from a teacher (Mayer, 2004). Empirical evidence and theoretical arguments do not support the usage of unassisted discovery learning (Alfieri et al., 2011; de Jong & van Joolingen, 1998; Kirschner et al., 2006; Klahr & Nigam, 2004; Mayer, 2004). Thus, guided forms of discovery learning are recommended, which include varying degrees of "…hints, direction, coaching, feedback, and/or modeling to keep the student on track" (Mayer, 2004, p.15).

Through guided discovery learning, instruction can be simultaneously informed by both constructivism and cognitive load theory. Both constructivist approaches and direct instruction may be employed by the same instructor, depending on the circumstance (Santrock, 2011; Schwartz & Bransford, 1998). Moreover, implementation of the constructivist pedagogies of inquiry and problem-based learning does need some scaffolding to manage cognitive load (see competing perspectives from Hmelo-Silver et al., 2007 and Kirschner et al., 2006). De Jong et al. (1996) provide an example of interspersing tasks as forms of guidance to reduce cognitive load in a discovery learning process. Implications for adopting a guided discovery learning approach for simulations are next considered.

2.2.2 Simulations as a form of guided discovery learning

Computer simulation for learning scientific principles has been cast as a discovery learning environment, where instructional supports are needed (de Jong et al., 1996; de Jong & van Joolingen, 1998; Swaak et al., 2004). De Jong and van Joolingen (1998) concluded that (1) across domains, results were mixed as to the degree to which simulations improved student learning over direct or expository instruction, and (2) lack of success in using a simulation for learning may be due to problems with unassisted discovery learning. When learning with simulation, pure discovery may suffer relative to other instructional techniques due to students having minimal necessary skills, such as planning which actions to take and not knowing how to create effective hypotheses and experiments (de Jong et al., 1996; de Jong & van Joolingen, 1998). Instructional tools and assistance are thus needed to manage cognitive load when exploring computer simulations (de Jong, 1991; de Jong & van Joolingen, 1998). Such guidance built into a simulation itself may include providing domain knowledge concurrently with the simulation, providing tasks or guiding questions, showing a progression of models, or establishing a structured framework to follow (de Jong & van Joolingen, 1998). See de Jong (1991) and de Jong and van Joolingen (1998) for in-depth review of discovery learning problems and additional support measures.

2.2.3 Extent of learning theoretic connections in the statistics education literature

There has been somewhat limited discussion of schema-building within the statistics education literature pertaining to simulations. Schemas have been used to frame conceptual development, including concepts of the sampling distribution (Lipson, 2002; Lipson et al., 2003) and statistical inference (Pfannkuch & Budgett, 2014). Notably in the

former case, student sampling distribution schemas benefitted from seeing both simulation-based and theoretical representations (Lipson, 2002). Such schemas may be developed in a stepwise fashion with interactive simulations, provided that students engage in particular behaviors (Lipson et al., 2003). Moreover, insufficient sampling distribution schema development at an early step may hamper the understanding of statistical testing at a later step.

References to using cognitive load theory have similarly been paltry. Budgett and Wild (2014) focused on how visualization balances germane load and extraneous elements, in an attempt to provide the best conditions for visually learning the randomization test. More extensively, Lipson et al. (2006) applied multimedia design principles (see Mayer, 2002) with emphasis on working memory constraints in an inference simulation. Finally, Chance and Rossman (2006) briefly referenced computer simulation usage throughout a course as a way to reduce cognitive load. These examples are limited in scope and detail, mainly focused on tying visualization to simulation.

On the other hand, employing simulations from a constructivist or knowledge-construction perspective in statistics courses is well documented (e.g., Ben-Zvi, 2000; delMas et al., 1999; Erickson, 2006; Lane & Peres, 2006; Mills, 2002; Wood, 2005). Lipson et al. (2003) described the purpose of constructivism as a means to building schemas, and succinctly, Erickson (2006) described simulation in statistics as a "constructivist's dream" (p. 3). Similarly, Lane and Peres (2006) proposed usage of computer simulations in a particular manner as a form of constructivist "active learning" that requires student engagement, instead of presenting simulations as passive demonstrations.

Simulations in statistics education have been framed as guided discovery in several areas (Lane & Peres, 2006; Lipson et al., 2006; Novak, 2014; Reaburn, 2014). Applying simulation throughout a whole course (Reaburn, 2014) and a whole online textbook (Lane & Peres, 2006) has been described as implementing guided discovery. However, passively observing a simulation has been classified as expository instruction, with discovery learning only activated by learner exploration and experimentation in a simulation (Novak, 2014). To reduce passivity and increase intellectual activity when interacting with simulations and to assist in guiding exploration, pedagogical strategies of predicting answers to questions about and outcomes from a simulation beforehand have been suggested (the "query-first" method; see Lane and Peres, (2006)) and implemented (delMas et al., 1999; Reaburn, 2014). As an example of explicit theory integration, Lipson et al. (2006) combined guided discovery learning with cognitive load theory, in an overarching cognitive model of multimedia design, to design a computer simulation that increased student statistical reasoning.

2.2.4 Linking learning theory to the characteristics of simulations

There are several aspects of simulations as they are used in statistics that may be argued as beneficial for learning from a theoretical standpoint. In some cases, the beneficial aspects may be explained through one or more characteristics of simulations (purpose, representation, and enactment).

First, the nature of simulations allows for breaking problems and concepts down into concrete steps and components (see examples of this in Cobb, 2007; Engel, 2010; Wood, 2005). This takes advantage of the enactment characteristic, or that a simulation is

defined by a procedure. This allows for crafting instruction with varied element interactivity and therefore customization of cognitive load.

Second, simulations allow for interaction with a physical, tactile process instead of formulae (Holcomb et al., 2010; Wood, 2005), revelation or transparency of phenomena otherwise mathematically or abstractly hidden (Holcomb et al., 2010; Wood, 2005), and matching the simulation process being represented (Cobb, 2007) to what students experience (Engel, 2010). These primarily relate to the representation and purpose characteristics of simulations, and presumably help to manage cognitive load by increasing the accessibility of the concept, phenomenon, or objective at hand. Furthermore, it is the non-mathematical and more accessible nature of simulations that may facilitate the opportunity for students to construct meaning with a simulation (Wood, 2005). That is, the more manageable cognitive load inherent to the constituents of a simulation may enable its constructivist benefits.

Third, limited prerequisite knowledge is needed to implement or explore simulations (Hesterberg, 1998; Holcomb et al., 2010). Hence, the cognitive load can be managed implicitly by novices.

Fourth, simulation as an inherently discovery learning environment when designed as such provides opportunity for the student construction of statistical concepts in an active manner (Mills, 2002; Wood, 2005), which can then be accompanied with instructional support (Lane & Peres, 2006; Lipson et al., 2006; Novak, 2014; Reaburn, 2014). For a given simulation, customizing the steps that students see or must take (i.e., adjusting the enactment and representation), provides an opportunity to customize guided instructional support.

Lastly, in the case of learning statistical inference, simulations may reduce the number of schemas that need to be managed at an early stage, in order to grasp a larger schema of inference. Since simulations as a whole may replace a multitude of situationally-dependent, distinct formulae (e.g., Cobb, 2007; Erickson, 2006; Holcomb et al., 2010), students may be able to better grasp the purpose of using simulations, compared to the purpose of selecting a formula for a given statistical question.

Taken together, the nature of and characteristics of simulations afford benefits based on the accessibility of its components (managing cognitive load and element interactivity) and the potential to offer exercises and environments that facilitate student constructions of knowledge (capitalizing on guided discovery). Ultimately, this depends on appropriate pedagogy and instructor decision-making, to support student schema-construction for targeted statistical concepts.

2.3 Review of quantitative empirical evidence for learning with simulation and resampling methods

The review of quantitative evidence is separated into two types of studies: (1) studies with designs focused on comparing multiple curricula across full classes, and (2) studies with other designs, such as comparing sections within a class, single activities within a class, and isolated interventions external to a class. To increase comparability of designs and outcomes, this set of studies was limited to those using a quantitative outcome to measure student learning.

Studies were identified with a semi-structured process. Titles and abstracts from all issues in the *Journal of Statistics Education* (https://amstat.tandfonline.com/toc/ujse20/current), *Statistics Education Research*

Journal (http://iase-web.org/Publications.php?p=SERJ_issues), and *Technology Innovations in Statistics Education* (https://escholarship.org/uc/uclastat_cts_tise) were reviewed for the terms of simulation, resampling, randomization, and any other method or description of an activity or process that could be classified as a simulation. Similarly, *Proceedings of the International Conference on Teaching Statistics* (ICOTS: http://iase-web.org/Conference_Proceedings.php) were reviewed in such manner. Extensive searches occurred stemming from reviewing previous work cited in relevant articles, as the relevant articles were acquired. This process resulted in a collection of works including both non-empirical and empirical study. In many cases, non-empirical work provided key history, theory, and simulation examples, and thus provided essential foundation for this review. Once the acquisition of works was exhausted, selection of papers for this review occurred via the criteria established above. This resulted in a total of 25 papers to be included, as of December 2018.

2.3.1 Attributes and results of studies comparing multiple curricula

Thirteen multi-class studies were identified that compared learning outcomes between entire classes, curricula, or a comparison group from a known assessment, using a clearly specified learning assessment. Sample sizes in study groups were typically in the hundreds, and occasionally on the order of tens or even up to thousands. For a detailed table of class comparison groups identified in each study, samples sizes, the learning assessment used for quantitative comparisons, and additional comments relevant to this set of studies, see Appendix A1: Attributes of studies comparing multiple curricula.

Study groups for comparison generally consisted of implemented curricula, courses, or textbooks; combinations of similar curricula, courses, or textbooks; or sets of scores from an assessment. For all studies, at least one study group was connected to a simulation-based course and another group was not. In many cases, studies employed groups with a clearly identified curriculum and specific textbooks. For example, Tintle et al. (2012) compared retention of learning outcomes from students in two different class types: one that used a preliminary version of the simulation-based *Introduction to Statistical Investigations* (ISI) textbook (Tintle et al., 2016) and one based on the traditional Agresti and Franklin (2008) textbook. The ISI curriculum was the most commonly identified curriculum in multi-class studies, though multiple versions of the ISI curriculum and textbook appeared across these. In other cases, a single study group may have consisted of students from one or more sections or courses known to use simulation-based inference (SBI) or traditional, non-simulation-based inference (non-SBI), without a clear textbook or version of a curriculum identified (e.g., VanderStoep et al., 2018). Thus, mixtures of textbooks and curricula may compose a single study group in many studies. Finally, a few studies employed a national sample of scores from some version of the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS) test (delMas et al., 2007) as a comparison group (e.g., Tintle et al., 2011).

The CAOS test or a modified version of it was used in all cases for comparing learning outcomes between groups, except in Beckman et al. (2017), which used a custom assessment for the study. Importantly, modified versions of CAOS introduced a number of changes, including one or more items intended to measure student learning in a simulation-based course (e.g., see description in Chance et al., 2016). Thus, use of

multiple versions of CAOS across studies may have introduced variation in the types of learning and comparisons being measured.

Usually, study results consisted of study group comparisons on learning assessment performance in some fashion. On the one hand, pre- to post-test scores were used for straightforward between-group comparisons in some cases (e.g., Tintle et al., 2011). On the other hand, some studies employed more complicated modeling of learning outcomes by curricular type, such as with logistic or mixed-effect models (Beckman et al., 2017; Hildreth et al., 2018). In all but three studies (Chance et al., 2016, 2018; Garfield et al., 2012), inferential comparative learning outcomes were provided.

Other differences in study focuses are worth noting. In two instances focus was given to retention following course completion (Mendoza & Roy, 2018; Tintle et al., 2012). In a different instance, cognitive transfer to new problems was the learning focus (Beckman et al., 2017). Another approach was that of Tintle et al. (2018) and VanderStoep et al. (2018), where students were stratified by multiple performance variables in order to evaluate learning outcome differences by student performance level. Lastly, Chance et al. (2016) put primary emphasis on whether learning gains could be clustered or explained by both instructor familiarity with simulation-based curricula and sets of student characteristics. In that case, the goal was not to compare outcomes to non-simulation-based curricula.

Based on nine of the ten multi-class studies that reported inferential outcomes at the full assessment level, groups using SBI either performed better than or no worse than non-SBI groups. Most study results either clearly favored (e.g., Tintle et al., 2014), or mostly favored SBI groups (e.g., Tintle et al., 2018; VanderStoep et al., 2018). However,

weak or unclear evidence favoring SBI over non-SBI curricula was also reported (e.g., Beckman et al., 2017; Mendoza & Roy, 2018). Notably, non-SBI groups were not clearly favored on any full assessment outcomes.

When considering empirical inferential results at the item level, similar evidence generally favoring SBI curricula is observed. When looking across items within a single assessment, sometimes SBI groups performed better than or about the same as non-SBI groups (Roy & Mcdonnel, 2018; VanderStoep et al., 2018), both better and worse than non-SBI groups (Hildreth et al., 2018; Tintle et al., 2014), and about the same as non-SBI groups or with weak evidence of differences favoring any group (Mendoza & Roy, 2018; Saputra & Couch, 2018). In summary, item outcomes seem to typically favor SBI groups than favor non-SBI groups, with occasional exception and depending on item and topic.

Only Chance et al., (2016) primarily focused on outcomes solely from simulation-based curricula, with limited, cautioned comparison of outcomes to instructors who taught non-simulation-based courses. Concerning other factors, conceptual gains were not well explained by clusters of student characteristics (e.g., age, GPA, attitudes, etc.) nor by clusters of teacher characteristics (e.g., tenure status, years teaching, etc.). Additionally, mixed-effect models (randomly varying by course section) yielded mixtures of student- and teacher-level variables and their interactions as having some explanatory value for conceptual gains (see Chance et al., 2016 for details).

2.3.2 Attributes and results of other types of studies with quantitative outcomes

Twelve studies were based on within-class group comparisons, multi-section group comparisons, comparisons over the span of a single class topic or unit, and isolated lab intervention studies. Study group sample sizes were under 100 in nearly all instances. For details on study design attributes for these studies, see Appendix A2: Attributes of other studies.

Comparison groups and study designs were highly variable and, in some cases, inconsistent across studies. Since known curricula were not the common form of a study group, author descriptions were used to identify study groups. In terms of general study types, two were external-to-class one-time interventions (Lane & Tang, 2000; Novak, 2014) while the rest could be classified as some form of a comparison of sections within a class (e.g., Maurer & Lock, 2016), iterations of a class (e.g., Reaburn, 2014), or some other form of a within-class or across-class comparison (e.g., Weir et al., 1991). Regarding the latter type, ambiguities in reporting led to ambiguities in group classifications. For example, Pablo and Chance (2018), used a version of ISI, which is a known curriculum. In contrast, while known curricula were typically not implemented, author descriptions were detailed enough to provide a clear label for a study group (e.g., Holcomb et al., 2010). However, less descriptive detail was provided in other studies, which limited understanding of what composed a study group and what students experienced (see Mills, 2004; Simon et al., 1976).

The variable nature of what study groups meant across these studies created obstacles. Most strikingly, the act of observing a simulation (with an audio script playing) was used as the simulation training method in Lane and Tang (2000). However, the non-simulation comparison group in both Weir et al. (1991) and Maurer and Lock (2016) also

used observation of simulations, while their interventions instead focused on different types of interactions with simulations. Thus, the comparability of study designs and groups and how they connect to study outcomes is reduced by such variability.

Additionally, not all studies employed non-simulation groups. For example, Francis et al. (2007) differentiated groups based on when students were exposed to simulation activity with respect to a lecture (before or after), while Novak (2014) compared two groups who completed a simulation activity that varied by the degree of integration of a storyline.

Notably, half of these studies involved some aspect of randomly assigning subjects or sections to study conditions. The most comprehensive example is from Maurer and Lock (2016), who randomized four cohorts (two per class) to simulation or non-simulation versions of the second-half of the same course, attempting to control for several logistical factors.

Since these studies mostly did not aim to compare full curricula, many instead emphasized one of several topic focal points, including the F-statistic in an ANOVA (e.g., Taylor & Doehler, 2015) and the Central Limit Theorem (e.g., Mills, 2004). Others included mixtures of topics, explained in varying levels of detail. For instance, Maurer and Lock (2016) provided course content details, whereas Simon et al. (1976) did not. Other studies provided paltry details on the topic of interest or the nature of how simulations were used, such as with Novak (2014) who only stated that the simulations were used for descriptive statistics.

Regarding learning assessments and reported outcomes, nearly all studies used class exam questions or other questions that were not connected to a clearly labeled

assessment. There tended to be reliance on only a select set of course exam items or other short sets of items that were apparently developed exclusively for the study or course. Two exceptions used items or an assessment of known origin (see Maurer & Lock, 2016; Pablo & Chance, 2018). Of additional importance is that several studies relied on quantitative scoring of qualitative questions (e.g., delMas et al., 1999; Reaburn, 2014). Moreover, the detail in reporting what was used as an assessment ranged from limited (e.g., Weir et al., 1991) to fully presented (e.g., Reaburn, 2014). In summary, assessments from these studies are best interpreted on an individual study basis.

Regarding learning findings, seven studies provided some form of a group comparison on an overall assessment or combined set of outcomes, with only four of these drawing comparisons between one or more simulation groups to one or more non-simulation groups. These results mostly suggest learning benefits of simulations above their respective non-simulation counterparts, or at least no difference in performance between those who were and were not exposed to simulations. Several outcomes exclusively favored the simulation group or groups (Simon et al., 1976; Weir et al., 1991), including a study with random assignment (Lane & Tang, 2000), while one result both favored simulation and showed no evidence for a difference between simulation and non-simulation groups, depending on the measure (Mills, 2004).

For item-level results, in the few comparisons of simulation to non-simulation groups (five studies), simulation generally was favored or showed no evidence for a difference (Lane & Tang, 2000; Weir et al., 1991), with mixed or limited evidence in other cases (Reaburn, 2014), which also included studies with random assignment (Maurer & Lock, 2016; Taylor & Doehler, 2015). Notably, a result of no difference was

desired in part of the results from Maurer and Lock (2016). Two items were based on traditional statistical theory, and thus no group differences indicated that the simulation group fared as well as the non-simulation group on non-simulation items.

Other study results focused on other types of learning outcomes and comparisons. As an example, delMas et al. (1999) demonstrated how updates to a simulation activity in response to disappointing student outcomes, led to performance increases on an assessment focused on visual judgments in a later class. Examples of other types of learning findings include: limited to no evidence for the effect of integrating a storyline in a simulation-based learning activity (Novak, 2014); limited to no evidence of learning benefits of using a flipped versus non-flipped classroom in a simulation course (Pablo & Chance, 2018); and no evidence of learning benefits from engaging in a manual simulation versus not, prior to engaging in a computer simulation (Holcomb et al., 2010).

2.3.3 Summary of results from quantitative empirical study

For the 25 studies in this review, there was notable variation in study types, reporting detail, usage of simulation, and how learning outcomes were tracked. Identifying this information was fairly straightforward for multi-class studies, which emphasized comparison between known curricula or types of curricula, and less straightforward for other types of studies.

Empirical findings broadly provide evidence for the learning benefits of simulation-based curricula and activities, compared to non-simulation-based curricula and activities. However, clear support for simulation-based courses has not been observed for all items, topic areas, curricula, and study approaches. Nonetheless, there is

no clear evidence that non-simulation approaches to learning consistently outperform simulation approaches.

2.3.4 Discussion of variations in reporting of curriculum implementation

There was notable variation in study implementation detail, particularly for the exact manner in which students engaged with simulations. In particular, some studies clearly explained the exact simulations used and how students interacted with them. As an example, Taylor and Doehler (2015) explained the activity that students experienced in detail, which focused on introducing the ANOVA F-statistic through randomization of cereal data. Moreover, in addition to including the exact topic and activity purpose, the concrete steps that students needed to take and the tools involved were provided. In other cases, the description of implementation was moderate to nonexistent. To illustrate: Mills (2004) described the activities of students in the simulation group only as "…[using] Excel to perform any experiments for the understanding of the abstract concepts…" (p. 21).

One issue underlying some of these reporting differences is the overall study design. For studies that involved numerous student interactions with simulations, such as with semester-long curricular studies, tracking and explaining every aspect of student engagement with each simulation was infeasible. This is especially apparent for studies comparing several curricula (e.g., Chance et al., 2018; Hildreth et al., 2018). Thus, by the nature of a smaller scope, smaller studies afford better opportunity to explain exactly how students were intended to interact with simulations.

To deal with the large number of possible intended interactions, a common approach in curricular-level studies was to explain the learning environment and

pedagogy. Most studies based on the *Introduction to Statistical Investigations* (ISI) curriculum were linked in this fashion, in that the desired pedagogy and course implementation were initially described in Tintle et al. (2011). Similarly, the Change Agents for the Teaching and Learning of Statistics (CATALST) curriculum and intended implementation were covered in detail by Garfield et al. (2012). Therefore, at minimum one could infer the general principles behind how students were intended to interact with simulations throughout a course. Nonetheless, as curricula have expanded to and been adjusted by different institutions, implementation differences have been expected and noted, with one factor being teacher experience with a given curriculum (see Beckman et al., 2017; Chance et al., 2016). Furthermore, predefined curricula may be knowingly adjusted and implemented differently than originally introduced. In Hildreth et al. (2018) they refer to implementing CATALST as one study group, except the authors acknowledged modifications, including substituting new activities for old ones, adjusting content, and changing the software partway through the study period. Aspects of ISI have similarly evolved. Therefore, consistency of the expected student interactions and study implementations even with well-defined curricula deserves careful interpretation.

A primary implication of reduced detail or clarity in reporting of implementation is the weakening of the connection between student interactions with simulations and observed learning outcomes. This is one slice of the potential conditionality of learning outcomes. That is, to what extent do the types of behaviors that students engage in and the concepts they are exposed to explain variation in learning outcomes from simulations? Exemplifying the types of possible connections between implementation and outcomes, de Jong and van Joolingen (1998) present empirically-supported

pedagogical practices to assist students' learning when using computer simulations in a scientific discovery learning environment. Learning theoretic perspectives and the characteristics of simulations may help illuminate this connection. However, without sufficient description of implementation, characterizing this connection and understanding how to use learning outcomes to inform future practice is mitigated.

2.3.5 Discussion of disentangling the learning effects of simulations from other factors

Another important issue related to study design is the ability to isolate the effects of simulations. In other words, to what extent can the effects of simulations on the observed learning outcomes be "disentangled" (Hildreth et al., 2018) from other factors? Given that the use of simulation is one quality embedded in a larger network of educational factors, disentangling learning outcome effects from content and pedagogy and establishing causality may be limited, particularly in observational studies of full-scale curricula (Hildreth et al., 2018; Maurer & Lock, 2016; Tintle et al., 2011). Hildreth et al. (2018) note study confounds of classroom, curriculum, course schedule, lecture versus activity-based pedagogy, the presence of a teaching assistant, and the fact that CATALST was designed as a terminal course while the other studied curricula were not. Hence, many pedagogical and other logistical factors may affect learning outcomes.

Controlling for other factors has included efforts to employ random assignment and the inclusion of covariates in models explaining learning outcomes. The most exhaustive example of random assignment is from Maurer and Lock (2016). Students were randomly assigned to one of four cohorts, which included instructors alternating who taught which cohorts throughout the course. Thus, cohorts primarily differed only by

curriculum (for the second half of the course), while instructor and other administration effects, such as the exact classroom or lab room, were theoretically balanced between cohorts. At the other end of the extreme, some studies conducted random assignment of students to experimental and control learning conditions for one-time intervention or lab studies with moderate to no connection to a course (see Lane & Tang, 2000; Mills, 2004; Novak, 2014). By having a more controlled experiment not occurring in the flow of a regular class, such studies may not reflect the reality of actual learning environments that students encounter. Thus, generalizability of the learning outcomes from such studies to students' experiences in courses may be limited.

Between these two extremes are examples where students or sections have been randomly assigned to learning conditions for small-scale classroom or activity studies that do not encompass the full duration of a course (e.g., Holcomb et al., 2010; Taylor & Doehler, 2015). In both instances, such small-scale class studies involved instructors being part of both control and experimental groups in some manner, offering additional examples of controlling for instructor type. Such studies present a balance between the difficulty of attempting to control for non-curricular variables while maintaining the reality of student experiences during a course, as opposed to a more controlled lab setting.

Especially for observational studies, attempts have been made to control for tracked variables in modeling learning outcomes. Beckman et al. (2017) controlled for institution (as a way to account for instructor familiarity with CATALST) and allowed for variation by course section in a mixed-effects model. Roy and Mcdonnel (2018) also used mixed-effect models in their analysis and accounted for sex of the teacher, course

duration, semester, type of incentive offered to student, sex of the student, and GPA. Concordantly, Hildreth et al. (2018) included variables of classroom, year, semester, and meeting day for the course. Examples of controlling for study factors outside of pedagogical factors include stratifying learning outcomes by other measures of student performance, such as ACT (e.g., Tintle et al., 2018; VanderStoep et al., 2018). Lastly, Chance et al. (2016) employed combinations of student- and teacher-level characteristics, such as student age, student attitudes, and teacher experience, as the primary goal of analysis. At a minimum, capturing and reporting key educational factors may afford some degree of control in explaining learning outcomes.

2.3.6 Discussion of assessment usage

Several other study design attributes are worth introducing to frame the study findings. The variety of assessments and topics considered across studies defines what counts as learning within each study and the degree to which study outcomes can be compared. Examples of the variety of assessments include entire test or assessment scores which included multiple concepts (e.g., Beckman et al., 2017; Pablo & Chance, 2018), topic subscales composed of multiple items (e.g., Hildreth et al., 2018; Saputra & Couch, 2018), or performance assessed from one or more items about a single concept or related cluster of concepts (e.g., Holcomb et al., 2010; Taylor & Doehler, 2015). In parallel fashion, studies that compared curricula naturally tackled multiple topics, while other studies may have only emphasized one or two topics. Thus, not all learning outcomes may be equivalently comparable. In multi-class studies, various forms of CAOS were typically used; therefore, studies that employed a version of this assessment are more comparable to each other than those that did not use this assessment. In several

cases, course tests or items from courses tests were used (e.g., Simon et al., 1976), or only one topic was emphasized with simulation (e.g., Reaburn, 2014). By addressing a small scope of topics, the applicability of such results is thus limited to those topics. Moreover, scoring was also a notable issue. In contrast to more developed assessments (e.g., CAOS, GOALS), some studies scored outcomes via author or research-team judgment of free-response prompts (e.g., Lane & Tang, 2000; Reaburn, 2014; Taylor & Doehler, 2015). The reporting of these scoring approaches varied in terms of the steps to taken to ensure scoring reliability. Therefore, reconciling the strategy and quality of scoring approaches may further reduce comparability.

Moreover, the issue of matching assessment to the simulation-based nature of courses has been raised. That is, to what extent does a given assessment measure what students are learning in a simulation-based course? As a primary example, CAOS was not designed specifically for simulation-based courses, and modifications to reflect what students are learning in simulation-based courses have been proposed (see Garfield et al., 2012; Tintle & VanderStoep, 2018).

In summary, evidence for learning outcomes are contingent on a given assessment to measure student learning (Maurer & Lock, 2016). Thus, understanding the learning effects from simulations at a quantitative level depends on assessment appropriate to learning with simulations. At the formal assessment level, CAOS has been implemented as a reliable measure of conceptual understanding; however, it was not designed to address simulation-based courses. This has led to several efforts to modify CAOS (see Chance et al., 2016) and design new assessments to reflect statistics courses that employ simulations (e.g., see Garfield et al., 2012; Tintle & VanderStoep, 2018). Additionally,

the sheer variety of informal assessments across studies indicate the large range of possible outcomes, understandings, and types of learning being assessed, with varying degrees of relation to aspects of simulations. In light of the evolving statistics classroom, attention should be put toward continuing to design assessment that matches what students are experiencing with simulations. Many learning outcomes at present may not fully reflect what students learned about simulations or how they might have linked their knowledge from simulations to specific concepts. Perhaps the variation in learning outcomes has largely been a function of variable assessments, or more strongly, perhaps some crucial aspects of learning from simulations have not yet been measured that would help inform the observed learning outcomes.

2.4 Misconceptions about simulations in statistics

Learning obstacles with simulation have been proposed and observed. These include difficulties with students transitioning between the sample level versus sampling distribution level and conflating the real world of the data with the hypothetical world of the simulation (Case & Jacobbe, 2018). Other potential learning obstacles include challenges unique to designing a process instead of selecting a formula (Erickson, 2006), student passivity in instances where simulations are shown instead of created and interacted with (Lane & Peres, 2006), and the fact that providing direction with what to pay attention to in a simulation does not guarantee additional learning benefits (delMas et al., 1999).

This subsection summarizes and extends the work on misconceptions specific to learning statistics with simulations. First, misconceptions that have been observed in students using simulations to work with null-hypothesis significance testing (NHST) are

synthesized. Then, these misconceptions are extended to how they might occur when using bootstrapping for statistical estimation. Second, the conceptual framework for statistical inference proposed by Case and Jacobbe (2018) is presented and extended to estimation with bootstrapping. Third, the broad misconception of conflating the hypothetical nature of the simulation with the real world is explored. This includes presenting different examples of this conflation and proposing three facets of simulation to better describe such conflations.

2.4.1 Types of misconceptions

Some work has covered the array of misconceptions students exhibit when working with statistical simulations (see Case & Jacobbe, 2018; Chance & McGaughey, 2014; Gould et al., 2010; Hodgson & Burke, 2000; Maxara & Biehler, 2007; Rossman & Chance, 2014; Saldanha & Thompson, 2002). Table 5 provides a list of misconceptions students have been observed exhibiting when working with simulation-based statistics courses or activities. Notably, the first twelve misconceptions were observed only when students were working with null-hypothesis significance testing (NHST) questions, primarily with a randomization test, and the last two misconceptions when working with probability questions. Thus, these misconceptions have not been observed when working with estimation questions. Accordingly, I proposed how each misconception might occur when answering estimation questions. Since bootstrapping is a common approach to estimation with simulation (e.g., Good, 2006; Hesterberg, 1998), I extended the description assuming that some form of bootstrapping is used. Hence, the extended descriptions provide an example of how a misconception might vary only with respect to bootstrapping when used to conduct statistical estimation. Lastly, it should be

acknowledged that some NHST questions may be answered with bootstrapping, and therefore the descriptions of each misconception as originally observed may vary even within NHST questions, if explored with different types of simulations.

The first seven misconceptions originated from two overarching themes of misconceptions from Case and Jacobbe (2018). These two themes are (1) the multilevel nature of inference, and (2) distinguishing the real world of the sample data from the hypothetical world of the simulated data. The first theme specifically encompasses distinguishing and transitioning among the true population, a single observed sample, and multiple simulated samples. This theme was built on the idea of students needing to balance a multi-tiered scheme, proposed by Saldanha and Thompson (2002). The second theme encompasses misconceptions tied to thinking simulation is acting in the real world or producing real-world data. Since multiple types of misconceptions were contained in each overarching theme, I restructured the results from Case and Jacobbe (2018) to propose a larger set of misconceptions captured by the two themes. Some of these first seven misconceptions were also observed by other authors. The remaining misconceptions were observed, proposed, or inspired from work by other authors, as noted.

Table 5 *Observed student misconceptions specific to working with simulation-based statistics*

Misconception	Observed description for NHST	Proposed Functioning in NHST vs. Estimation (with bootstrapping)	Description for Estimation (with bootstrapping)
1. Number[a,b]	Confusing sample size vs. number of trials	Same	Confusing the sample size vs. number of bootstrap samples
2. Unit[a,c]	Not understanding the individual units in the simulated sampling distribution	Same	Not understanding the individual units in the simulated bootstrap distribution
3. Level[a,b]	Not transitioning to the sampling distribution level for inference	Same	Not transitioning to the bootstrap distribution level for inference
4. Method[a]	Mixing simulation- and theory-based methods within the same task	Not applicable	Combining bootstrapping with an aspect of traditional methods may be intended, such as bootstrapping to estimate the standard error for use in a traditional confidence interval formula
5. Replication[a,c,d,e]	Treating simulation as a replication study or sample	Same	A bootstrap simulation may be similarly interpreted as replicating real sample data
6. Center[a,f]	Misusing the center of the simulated sampling distribution	May apply differently	The simulated bootstrap distribution center may be less of a focus compared to NHST and should be near the sample statistic
7. Discredit[a]	Discrediting an observed sample as a fluke or outlier with respect to the simulated sampling distribution	May apply differently	The observed sample is not used to evaluate the tenability of a model, but that does not preclude the possibility of disregarding the sample data for other reasons,

Misconception	Observed description for NHST	Proposed Functioning in NHST vs. Estimation (with bootstrapping)	Description for Estimation (with bootstrapping)
			such as non-random sampling
8. Purpose[d]	Not recognizing the purpose of randomness or of the null hypothesis in simulation	May apply differently	Not recognizing the purpose of assuming the original sample represents the whole population or of bootstrapping or the role of randomness in this method
9. Technology[c]	Identifying any usage of technology as a simulation	Same	Identifying any usage of technology as a simulation
10. Variability-neglect[c]	Fixating on the center and shape of the simulated sampling distribution, in neglect of the variability	Same	Fixating on the center and shape of the simulated bootstrap distribution, in neglect of the variability
11. Parameter[c,d]	Not understanding what the population parameter is	Same	Not understanding what the population parameter is
12. Transfer[d]	Difficulty in designing simulations for new situations	Same	Difficulty designing a bootstrap procedure for a new situation
13. 50/50[d]	Difficulty in moving from two-outcome 50/50 proportion models to models of other proportions	May apply differently	An equally-likely two-outcome null hypothesis does not have an equivalent in bootstrapping; there may not be unique difficulty in abandoning a .5 proportion bootstrap compared to abandoning any other proportion

Misconception	Observed description for NHST	Proposed Functioning in NHST vs. Estimation (with bootstrapping)	Description for Estimation (with bootstrapping)
14. Transform[g]	Difficulty transferring a defined experiment to software	May apply differently	Given different simulation types used for estimation, there may be different kinds of difficulties with transfer to software
15. Process[g]	Other problems understanding the dynamics of the simulation not captured by other misconceptions	May apply differently	Given different simulation types used for estimation, there may be different kinds of difficulties with understanding the simulation dynamics

Notes. NHST = Null Hypothesis Significance Testing. Misconception sources for NHST: [a]Case and Jacobbe (2018), [b]Saldanha and Thompson (2002), [c]Rossman and Chance (2014), [d]Chance and McGaughey (2014), [e]Hodgson and Burke (2000), [f]Gould et al. (2010), [g]Maxara and Biehler (2007).

All misconceptions are expected to manifest in NHST and estimation with some similarity, except for the Method misconception. In NHST, a Method misconception requires inappropriately mixing simulation and traditional methods. However, in estimation with bootstrapping, it may indeed be intentional to mix methods, such as estimating a standard error with bootstrapping while choosing a standard error multiplier using a t-distribution.

The Center, Discredit, Purpose, 50/50, Transform, and Process misconceptions may be applicable to estimation but only in a substantively different manner compared to NHST. Since the purpose of estimation with bootstrapping is to mimic and emphasize sampling variability, the center of the simulated bootstrap distribution may be of reduced emphasis. Additionally, since the center of the simulated bootstrap distribution should be

similar to the original sample statistic, this may be a less notable comparison, in contrast to reconciling why the center of the simulated sampling distribution may be different than the original sample in NHST. However, an issue novel to estimation with bootstrapping is that students may think the center of the simulated bootstrap distribution is "more accurate" and disregard the sample estimate, which is related to the Discredit misconception. While evaluating the tenability of the model is not a step in bootstrap estimation, the sample of data may be discredited for other reasons.

The Purpose misconception as proposed by Chance and McGaughey (2014) pertains to the purpose of a null hypothesis and randomness in NHST, which have rough equivalents in estimation with bootstrapping. For the null hypothesis, that relationship exists at the "first level" according to the Case and Jacobbe (2018) inference framework, shown below. This level is where the hypothetical relationship in NHST or the real relationship in the population exists. In estimation with bootstrapping, the equivalent Purpose misconception may be not understanding the purpose of treating the sample as representing the population or of what randomness provides in mimicking random sampling.

The 50/50 model in NHST pertains to moving from equally likely two-outcome situations to outcomes with uneven probabilities or more than two outcomes. However, in estimation, there may be no special cognitive aspect when moving from a .5 proportion to another proportion. Hence, transferring between estimation situations may be a different flavor of misconception, potentially already captured under the Transfer misconception.

Lastly, given the broad nature of the Transfer and Process misconceptions, each presumably operates in a variety of different ways in answering estimation questions.

Since different types of simulations are employed for estimation questions, the problems when defining the simulation abstractly (Process) and implementing it in software (Transfer) are likely different compared to NHST.

Due to the fundamental differences between the operationalization of estimation with bootstrapping versus other statistical simulations, there may be important misconceptions unique to estimation not covered in Table 5. For example, producing a single bootstrap sample requires sampling with replacement the same number of times as the original sample size, which is a step and concept not present in a two-group randomization test. That is, misunderstanding the difference between the sample size versus the number of trials in a randomization simulation may be a wholly different misunderstanding than knowing that the bootstrap sample size needs to match the original sample size.

2.4.2 Thematic organization of statistical inference

To organize student misconceptions, Case and Jacobbe (2018) proposed a framework for simulation-based inference. Since this framework was constructed only under situations with the randomization test for NHST, I expanded it to include estimation questions specifically when bootstrapping is the simulation type, as shown in Table 6.

Case and Jacobbe (2018) organized the framework along two dimensions: a constituent "level" within statistical inference and the "world" where the inference levels exist. At the first level, there is a whole population or true relationship out in the real world (R1). The analog for NHST is the population or relationship that is described by the null hypothesis in the hypothetical world (H-S1). Extending the first level to

estimation is somewhat unclear. Estimation with bootstrapping does not propose an exact model or relationship for testing. Nevertheless, bootstrapping is built on the proposition that the sample is a proxy for the whole population (Hesterberg, 2015; Lock et al., 2013). Thus, the estimation analog to a real or hypothetical population is treating the sample as a representation of the whole population (H-E1). As this representation is a "pseudo-population" (Mooney, 1997), it may also be regarded as a kind of hypothetical world. The second level consists of single-sample distributions: the real-world original sample (R2), a hypothetical simulated sample (H-S2), and by extension, a bootstrapped sample (H-E2). Finally, the third level involves the sampling distribution: the real-world sampling distribution generated from actual replication (R3), the simulated sampling distribution generated from simulation (H-S3), and again by extension, the simulated bootstrap distribution (H-E3).

Table 6 *Framework for organizing simulation-based inference*

Real-world	Hypothetical (significance testing)	Hypothetical (estimation with bootstrapping)
Whole population or true relationship (R1)	Hypothetical population or relationship (H-S1)	Using the observed sample as a representation of the whole population (H-E1)
Observed sample data (R2)	Distribution of one simulated sample (H-S2)	Distribution of one bootstrap sample (H-E2)
Distribution of statistics produced through replication (R3)	Distribution of statistics produced through simulation (H-S3)	Distribution of statistics produced through bootstrap simulation (H-E3)

Note. The first two columns are from Case and Jacobbe (2018).

2.4.3 Conflating the hypothetical nature of simulation with the real world

Study of student understanding of the hypothetical nature of simulation versus the real world of the data, and the consequences of such understanding or lack thereof, are not apparent in the literature. In contrast, student understanding of inference "level", particularly with NHST has been studied before (e.g., Saldanha & Thompson, 2002). To begin to understand the role of real-world conflation in the statistics classroom, examples of observed student misconceptions that appear to arise from this conflation are next considered.

The Replication, Center, and Discredit misconceptions observed by Case and Jacobbe (2018) each involve some aspect of conflating the simulation with the real world. The Replication misconception involves viewing the simulation as generating real data, which invokes mixing up hypothetical and real worlds. Similarly, the Center misconception involves some form of misinterpreting the center of a simulated sampling or bootstrap distribution by thinking that simulation is occurring in the real world and thus the center of the distribution is the true value in the real world. Such a mix-up of worlds may further lead one to discredit the real sample data, as not fitting in with the simulation (i.e., the Discredit misconception).

Observed examples of misconceptions suggesting real-world conflation may be categorized in multiple ways. Table 7 presents different types of examples of misunderstandings that may source from real-world conflation. These examples are then categorized in two ways. The first way is by the type of misconception that the example seems to be demonstrating. These misconceptions were drawn from Table 5. The second way is by the main facets of the simulation that appear to be the subject of misinterpretation. I proposed these facets, which are next discussed.

Based on a synthesis of the observed examples suggesting real-world conflation, there are three facets of simulation that appear to be invoked. These facets will be labeled "Process", "Product", and "Panacea". These facets refer to the part of the simulation that is the focus of the real-world conflation. The Process facet refers to conflating the hypothetical processes of the simulation as being a real-world process. The Product facet refers to conflation of the hypothetical product of the simulation with a real-world product. The Panacea facet refers to giving a hypothetical simulation powers that fix or affect aspects of the real-world data or results. Each example of conflation seems to primarily connect to one or two of the three facets.

Table 7 *Types of observed misconceptions suggesting real-world conflation*

Example Source	Type	Related Misconception	Main of Facet(s) of Misunderstanding
Hodgson and Burke (2000)	Simulation generates samples to describe the true population parameter from a known distribution	Replication	Process, Product
Rossman and Chance (2014)	Simulation replicates prior research to strengthen findings	Replication	Process, Panacea
Case and Jacobbe (2018)	Simulation can correct flaws in the original study supplying the data	Replication	Panacea
Gould et al. (2010)	A simulated distribution is a real-world distribution, leading to misinterpreting the center as supporting the null hypothesis in NHST	Center	Product
Case and Jacobbe (2018)	Using the center of the simulated distribution as the starting point to compute a p-value in NHST	Center	Product
Case and Jacobbe (2018)	A low p-value means the observed data are unlikely and thus should be discarded	Discredit	Process, Product

Note. Example types are from the stated sources. The proposed underlying misunderstanding is a mixture of my own proposition combined with description by the source authors.

The conflation example from Hodgson and Burke (2000) was assigned both the Process and Product facets. It was assigned the Process facet because the act of generating a sample to describe the true population parameter suggests a real-world process. It was assigned the Product facet because the samples and parameter that are implicated are real-world outcomes. As Hodgson and Burke (2000) state: "…the students

appeared to view the purpose of gathering large numbers of samples to be simply a real-world strategy for finding a population parameter." (p. 94).

The second example was assigned the Process and Panacea facets. From Rossman and Chance (2014): "Some mistakenly believe that simulation aims to provide replication of the research study, in order to strengthen the findings through replication." (p. 218). As described, providing replication sounds like a real-world process. Regardless of whether such students believe this replication is a real-world process, further referring to "strengthen the findings" suggests additional real-world validity or power is attributed to simulation that is not actually occurring. The Product facet was not assigned, as the focus of this example was not on a specific product from the simulation.

Thinking that simulation can correct flaws in the data was assigned the Panacea facet. As reported by Case and Jacobbe (2018, p. 22-23), one student in their study thought that a randomization simulation could eliminate confounding factors, because it was executed "over time", while a theory-based statistical test could not. Thinking that a simulation can correct flaws in the original data is tantamount to imbuing the simulation with special powers to affect the real world that it does not have.

The conflation example from Gould et al. (2010) was assigned the Product facet. As they observed: "A common misconception students reveal in the labs is that the null distribution of the test statistic is seen as the "real" distribution, and students reason that because the distribution is centered at 0, the null hypothesis cannot be rejected." (p. 4). This illustrates that the simulated distribution is directly interpreted as a real-world outcome. This conflation then leads to misusing the center of the simulated distribution. Similar to the prior example, Case and Jacobbe (2018) present another type of misusing

the center, also assigned the Product facet. They discuss a student in their study using the center of the simulated distribution to compute a p-value. One explanation for this choice is that the student is connecting the distribution to a real-world product.

Lastly, discarding the real data due to a low p-value was assigned the Product and Process facets. Case and Jacobbe (2018) observed a student viewing the simulation as replication and thus seeing an unlikely result as unlikely in reality. (Notably, this misinterpretation was corrected in one student once they encountered a conflicting result from the accompanying theory-based approach on the same problem. (p. 23-24).) This was assigned the Product facet, because discarding the real data suggests the simulated data are given real-world prioritization as outcomes. This was assigned the Process facet, as Case and Jacobbe describe this as a conflation of simulation and replication. Implicating replication suggests that the simulation is being connected to a real-world process.

In summary, each of the three facets is intended to capture a different aspect of conflating a hypothetical simulation with reality. However, distinguishing which facets are invoked for a given misconception is subjective. Nothing precludes all three facets from being involved in a given conflation. For example, a student who thinks that simulation can correct flaws in the study may be viewing the simulation as a real-world process, producing a real-word product, and offering a power to affect the real-world result. Determining the facets of simulation that a student may be conflating with the real world is dependent on the observed language the student uses to describe their thinking. While a student may express a conflation suggesting that all three facets are viewed as

real, different types of conflations may primarily only focus on or use explicitly language that reflects a subset of facets.

Adding further complication, students may shift between hypothetical and real worlds. Case and Jacobbe (2018) describe a student that correctly modeled a null hypothesis with a coin but later made a conflation error. They argue this meant the student correctly saw the simulation as a hypothetical world but then shifted to a real-world interpretation of simulation when making the error. Citing Johnson and Lesh (2003) and Lesh and Doerr (2003), they explain: "These inconsistent interpretations of a system may be associated with the use of representational media that emphasize different aspects of the underlying systems." (p. 25). Hence, students may invoke different conflation facets at different phases of an analysis. As Case and Jacobbe conclude: "…this study found that students' coordination of real-world and hypothetical perspectives was "unstable"." (p. 25).

2.5 Problem Statement

Little is known about the extent and nature of students' real-world conflation in simulation-based statistics courses. Even less is known about how harboring such conflations may affect understanding other statistical ideas. Finally, the design of assessments to evaluate student understanding of simulation-based content is still expanding. Therefore, student thinking about the hypothetical nature of simulations should be elicited in an open-ended manner. For a complete picture, such thinking should be elicited in multiple situations about multiple parts of the statistical analysis process; therefore, this is what I will do. This will provide observational evidence about the scope and variety of misconceptions related to real-world conflations. This evidence will further

help to define the content of future assessments and offer recommendations to statistics educators for the types of conflation-based misconceptions to expect in their classes.

Chapter 3: Methods

This study aimed to answer the following research questions: (1) *To what extent are there quantitative differences in student understanding of the hypothetical nature of simulations when working with null-hypothesis significance testing vs. estimation?*, and (2) *What typical themes emerge that indicate students are conflating the hypothetical nature of simulations with the real world?*

The remainder of this chapter describes the methods that were used to answer the research questions. First, an overview of the study with a timeline of major events is presented. Second, how the initial instrument was developed is described. This includes the initial blueprint. Third, the development and execution of the Think-aloud interviews and consequent changes to the instrument and blueprint are detailed. Fourth, the development and execution of the field test is described. Fifth, the data analysis is detailed. This includes descriptions of the scoring rubric development, quantitative scoring process, qualitative classification process, and analysis steps for answering both research questions.

3.1 Study overview

The purpose of this study was to identify the presence of and classify introductory statistics students' real-world interpretations of hypothetical simulations. Specifically, the goal was to compare interpretations of a randomization test used for a null-hypothesis significance test (NHST) to a bootstrapping simulation used for an estimation task. The Simulation Understanding in Statistical Inference and Estimation (SUSIE) instrument was created to elicit and collect data on such interpretations. The instrument focused on three different facets of the hypothetical nature of simulations (Process, Product, and

Panacea) and consisted of two sections: one data analysis situation focused on NHST with the randomization test and one data analysis situation focused on estimation with bootstrapping. Each section contained a description of the situation and nine constructed-response items.

The study took place from October 2019 through June 2021. See Table 8 for a timeline of the main steps of the study. The primary method for collecting data was web-based administration through Qualtrics. The main study phases consisted of developing the instrument, piloting the instrument with Think-aloud interviews, collecting primary data via Qualtrics for the field test, and analyzing results. Initial instrument development consisted of drafting descriptions of two statistical tasks, twenty constructed-response items, and an instrument blueprint. These draft materials were based on student misconceptions from the literature, the three facets of conflation, and my classroom teaching experience. The initial version of the instrument was finalized after informal discussion and feedback from my advisors and colleagues with expertise in statistics education. The initial instrument version was piloted in web-based Think-aloud interviews conducted with college students with a variety of statistics expertise. The instrument was iteratively updated through the piloting process until a final eighteen-item version was created, which was used in the field test.

The final version of the instrument was implemented in Qualtrics and responses from participants were collected for sixteen days at the end of the fall 2020 semester. Participant recruitment for the field test involved recruiting college students from introductory statistics courses at the University of Minnesota that used a simulation-based curriculum. To quantitatively score the accuracy of responses, I created a scoring rubric,

which was edited through a peer-assisted validation process. I made additional changes while scoring responses. The first round of data review consisted of quantitatively scoring responses and assigning qualitative codes to incorrect responses. The second round of data review involved assigning new qualitative codes to incorrect responses. The third and final round of data review consisted of assigning new qualitative codes only to incorrect responses that used language suggesting a conflation of the simulation with the real-world. Quantitative data analysis consisted of modeling and evaluating the total score on the instrument and the difference in scores between the NHST and estimation sections. A brief reliability analysis was also conducted. Qualitative data analysis consisted of aggregating responses that used language suggesting a real-world conflation, evaluating the prevalence of such conflation-language across items, and identifying the typical themes of responses that used such conflation-language.

Table 8 *Timeline of main steps to complete the study*

Step	Timeline
Define target content areas for instrument	October 2019 – January 2020
Draft initial blueprint, situations, and items for SUSIE	January 2020 – February 2020
Revise and create final version of blueprint and instrument for Think-aloud interviews; draft study protocols for Think-aloud interviews and fielding testing	February 2020 – May 2020
IRB approval	May 27, 2020
Finalize study protocols for Think-aloud interviews	May 2020 – June 2020
Execute test-runs of Think-aloud interviews	June 12 and June 13, 2020

Step	Timeline
Send email request to instructors to recruit students for Think-aloud interviews	June 15, 2020
Execute twelve Think-aloud interviews while iteratively revising the instrument	June 18, 2020 – August 31, 2020
Contact instructors for field test recruitment	August 31, 2020
Create final draft of SUSIE and extra credit survey in Qualtrics	September 29, 2020
Confirm final field test logistics with instructors	October 19, 2020 – November 14, 2020
Create initial draft of scoring rubric	November 3, 2020
Recruit participants for field test	November 19, 2020 – December 9, 2020
Open SUSIE for field test data collection in Qualtrics	December 1, 2020
Close SUSIE for field test data collection in Qualtrics	December 17, 2020
Send list of participants to respective instructors for extra credit incentive	December 17, 2020
Complete peer validation of scoring rubric and update rubric	January 4, 2021 – January 21, 2021
Execute quantitative scoring and first round of qualitative coding of participant responses	January 28, 2021 – March 25, 2021
Update final scoring rubric	April 3, 2021
Execute second round of qualitative coding	April 5, 2021 – April 26, 2021
Execute data analysis and synthesize results	April 26, 2021 – June 21, 2021

3.2 Initial development of SUSIE

The initial instrument was developed for the pilot Think-aloud interviews. The objective was to create an instrument with only constructed-response items focused on different aspects of the hypothetical nature of simulations. Creating this instrument involved drawing on the literature of observed misconceptions and extending the coverage of possible misconception content. This involved employing the three facets of Process, Product, and Panacea proposed earlier. The instrument contexts, items, and blueprint were drafted in a non-linear manner, often simultaneously, as described below.

3.2.1 Intended population

The intended population is college students who have completed an introductory simulation-based statistics course. Selection of this population was based on the need to further evaluate learning outcomes and update assessment in the modern introductory statistics course, as discussed in Chapter 2. Choosing this population informed the scope of the target content, blueprint, and item-writing.

3.2.2 Instrument format

The format of the instrument was determined by the objective to elicit students' conceptual interpretations of different aspects of statistical simulations that might involve a real-world conflation. The format of the instrument consisted of three key properties: use of only constructed-response items, presenting situations that illustrated the entire process of a statistical analysis, and presenting one situation that employed null-hypothesis significance testing (NHST) and one situation that used estimation.

Regarding item format, the instrument employed constructed-response (open-ended) items in contrast to selected-response (closed-ended) items. The constructed-response format was chosen due to the limited understanding of real-world conflation.

Use of items to elicit open-ended responses allowed for novel interpretations of simulations to emerge that might not have been captured with closed-ended item formats. This item format supported the second research question of identifying different examples and themes of real-world conflation.

Also because of limited understanding about real-world conflation, the instrument was designed to present descriptions of full statistical analysis processes. By presenting descriptions of the statistical analysis process, this allowed the content of the items to focus on conceptual interpretation of each step of the process. The intent of the instrument was not to have participants perform computations or execute an analysis. Instead, to explore the maximum scope of where real-world conflations may exist for students, the objective was to elicit student thinking on all parts of the statistical analysis process that could be reasonably described in a written test setting.

Finally, misconceptions about simulation in the literature have only been discussed from the standpoint of NHST, typically with the randomization test. Therefore, two contexts for motivating the items were created. One context involved answering a statistical research question with NHST using the randomization test, and the second context involved answering a statistical research question with estimation using a bootstrap procedure. Using two different contexts allowed for the responses to these two types of simulations to be compared. Presenting a bootstrap simulation further allowed the understanding about real-world conflation to be broadened to a situation other than a randomization test or NHST. The instrument content, items, and blueprint were informed by this format and are discussed next.

3.2.3 Instrument contexts, items, and initial blueprint

Drafting the instrument situational contexts, statistical analysis descriptions, items, and blueprint occurred in a non-linear and often simultaneous manner. ("Blueprint" refers to an organized presentation of the items with a summary of the key features about each item.) The content considered for the instrument was defined by the real-world conflation examples in the literature and the two contexts each containing descriptions of a full statistical analysis process. While drafting the instrument contexts and statistical analysis descriptions, new types of possible real-world conflations were conjectured and subsequently integrated into the drafts of items. Through iteratively drafting and updating the contexts and items, the blueprint for the instrument was also drafted and updated. Finalizing the blueprint then assisted in finalizing the instrument contexts and items. Throughout the instrument development process, different iterations of the contexts, items, and blueprint were reviewed and discussed with my faculty advisors and colleagues in statistics education. The initial complete draft of the instrument consisted of descriptions of two statistical research contexts with accompanying research questions, two descriptions of complete statistical analysis processes, and twenty constructed-response items integrated throughout the analysis process descriptions. For the initial complete draft of the instrument, see Appendix B1: Initial SUSIE versions for Think-aloud interviews. The components of the instrument and the initial blueprint are next described in turn.

Two instrument contexts were drafted. The first context involved modifying the fish oil study originally published by Knapp and Fitzgerald (1989). This study has been modified for textbook and study usage before (e.g., see Ramsey & Schafer, 2012, p. 245 or Tabor et al., 2012, p.23-24). The objective for the first instrument context was to

present a research question that needed to be answered with NHST and use a randomization test. The premise of this context is that fourteen male subjects were randomly assigned to consume fish oil or another oil in a four-week diet. The change in blood pressure was then measured (post-study minus pre-study), to evaluate if oil consumption has a relationship with blood pressure. The version used for the instrument was initially based on the version reported in Case (2016) but was modified in several places. First, two characters called “Researcher 1” and “Researcher 2” were added. These characters would become the basis for the statements proposed in many items and serve to drive the description of how the statistical analysis for the context was carried out. Second, a research question for the context was integrated and proposed by the characters, as stated: “Is there a difference in the average blood pressure reduction between the fish oil group and soybean oil group?” By asking, “Is there a difference…” this research question intended to cue study participants that NHST would be used in the analysis. Third, the sample data (the fourteen values of blood pressure change) were shown on vertically stacked dot plots, conditioned by oil group. Fourth, computation of the sample average difference in the reduction in blood pressure (of 7.7, favoring the fish oil group) was shown. Lastly, other wording adjustments were made to increase the clarity of the context. The data for this context were the fourteen blood pressure values reported in Case (2016) and elsewhere.

The second context was my own creation, which focused on delays in airplane departures. The objective was to present a situation that necessitated an estimation research question. To parallel the format of the first context, two new characters were presented, “Friend 1” and “Friend 2”. The premise of this context is that these two friends

want to know the average delay for recent airplane departures from Minneapolis-St. Paul (MSP) airport. While the first context involved an experimental study with random assignment, this context employed the random sampling of 150 departure delay times from a specific database. To simplify the context, only Delta Airlines flights from 2019 were considered. Similar to the first context a research question was proposed, but this research question was intended to cue an estimation procedure for study participants: "What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?" Also similar to the first context, a plot of the sample data was provided and the sample average delay (of 6.35 minutes) was displayed. To acquire the data for the context, data for all Delta flight departures in 2019 were downloaded from a Bureau of Transportation Statistics database (https://www.transtats.bts.gov/ONTIME/Departures.aspx). From the population of all 63,049 departures, different sets of 150 departure values were randomly sampled, until a set of 150 values resulted in a sample average similar to the population average and a sample distribution similar to the population distribution. This was done to provide study participants with a realistic sense of the population.

Descriptions of full statistical analysis processes that followed the introduction of each context were drafted. Within each instrument section (i.e., for both contexts), the objective was to describe all major steps that needed to be taken, in order to answer the proposed research question at the introduction of the given context. The major steps presented in each instrument section were a description of the simulation that was executed for data analysis, a plot of the simulated sampling distribution, a description of the relevant information and computation from the simulated sampling distribution,

answering the research question, and considering several post-study issues. To maintain continuity and logical flow of the analysis, the steps were described from the standpoint of the actions of the two characters (i.e., the two researchers for the NHST context or the two friends for the estimation context). To support the parallelism between instrument sections, the major steps were written in the same order and in as similar a format as possible between each instrument sections.

Each simulation description was presented as a set a of steps used to accomplish a purpose specific to the given research question. For the NHST section the stated purpose was "[t]o evaluate whether the difference of 7.7 was simply due to random chance", and for the estimation section the stated purpose was "[t]o estimate the uncertainty in the average of 6.35". For the NHST section the simulation was called a "computer simulation", while for the estimation section it was called a "bootstrap simulation". The steps of each simulation were presented as a generic algorithm, not specific to any software. The same sentence was provided at the end of the simulation description, in each instrument section: "The computer repeated the process above for a large number of trials."

Following the simulation description in each section, items were created and integrated throughout the remainder of each statistical analysis description. Typically, the description of an analysis step served as the introductory text for the item. Thus, much of the analysis description was written simultaneously with one or more items. The objective for item-writing was to first integrate previous examples of real-world conflation among the analysis descriptions, where possible, and then draft items that would potentially elicit novel examples of real-world conflation. The previous examples

were directly drawn from those introduced earlier in Table 7. To ensure sufficient content coverage, items were proposed such that participants would have to provide conceptual interpretations on many different parts of the analysis process. Additionally, the content of items was drafted such that each of the three facets of conflating simulation with reality (Process, Product, and Panacea) were sufficiently targeted.

An initial set of twenty items, comprised the set of items to be evaluated in Think-aloud interviews. An initial classification of items served as the initial blueprint, as shown in Table 9. The first ten items were for the NHST section (indicated by the leading "iN") and the second ten items were for the estimation section (indicated by the leading "iE"). The "i" refers to the fact that these were the initial item designations and that the numbering of items changed by the final blueprint used in the field test. A description of what each item focused on is described by the item content focus. Each item was assigned what the primary facet(s) of real-world conflation would be, were there to be a conflation in a participant's response to the item. Given the subjective nature of the three facets and given the novel nature of this topic, the assigned facet(s) were a logical judgement based on the item content. It was assumed that items may elicit any combination of or all three facets in a participant's response. The underlying misconception(s) that informed the creation of the item were also assigned.

Table 9 *Initial item blueprint*

Initial Item Designation	Item Content Focus	Facets (misconceptions)
iN1	Null hypothesis and assumption of simulation	Process (Purpose)
iN2	How assumption of simulation helps answer research question	Process or Panacea (Purpose or Replication)
iN3	Meaning of distribution center around 0	Product (Center)
iN4	Using distribution to calculate p-value	Product (Center)
iN5	What a low p-value means	Product (Discredit)
iN6 (versions)	Value of more trials	Process or Panacea (Purpose or Replication)
iN7	Simulation and confounding variables	Panacea (Replication)
iN8	Simulation and sample size	Panacea or Process (Replication)
iN9	Simulation and unreliable data measurement	Panacea (Replication)
iN10	How to replicate study	Process (Replication)
iE1	Assumption of sample as stand-in for population	Process (Purpose)
iE2	How sample as stand-in for population helps answer research question	Process or Panacea (Purpose or Replication)
iE3	What to use for sample estimate in interval calculation	Product (Discredit)
iE4	Need for interval to answer research question	Product (Center)
iE5	Deviation of sample estimate from center of distribution	Product (Discredit or Center)
iE6 (versions)	Value of more trials	Process or Panacea (Purpose or Replication)
iE7	Simulation and sampling bias	Panacea (Replication)
iE8	Simulation and sample size	Panacea or Process (Replication)
iE9	Simulation and unreliable data measurement	Panacea (Replication)
iE10	How to replicate study	Process (Replication)

To create to parallel sections, items were paired between sections, based on their item content. The first item pair, iN1/iE1, asked about what was assumed to be true about each respective simulation. The second item pair, iN2/iE2, was written to directly follow

from iN1/iE1. Each of iN2/iE2 asked how the assumption from the respective iN1/iE1 contributed to answering the respective research question.

Items iN3 and iE5 were paired, given that each item addressed an erroneous interpretation of the center of each simulated distribution. The items appeared in different relative item orders, due to the analysis differences with NHST using a p-value and estimation using an interval. A character in iN3 proposed that a simulated distribution being centered around 0 supports the null hypothesis, an erroneous interpretation. For item iE5, an erroneous statement is offered that suggests the sample mean and data cannot be trusted due to the difference in the bootstrap distribution and sample means.

The pair of items iN4 and iE3 both addressed a calculation aspect of answering the respective research questions. Item iN4 had the characters offer competing ways to compute the p-value, with the erroneous computation starting at the center of the simulated distribution. Item iE3 had the characters offer competing ways to compute the confidence interval, with the erroneous computation using the mean of the bootstrap distribution instead of the sample mean.

Both of iN5 and iE4 focused on interpreting results to answer the respective research questions. For item iN5, two competing interpretations of the low p-value are offered, wherein the erroneous interpretation argues for the sample mean not being supported by the data. For item iE4, a character proposed that the average of the simulated distribution can be reported in lieu of a confidence interval, also an erroneous interpretation.

For items iN6/iE6, two wording variations to be tested in the interviews were created. The intent was to create two items to elicit how participants perceived the

running of trials, such as whether that amounted to replicating new data. “Wording A” presented four images of simulated distributions from four different numbers of trials and asked about how differences in the images might be related to the differences in the number of trials. “Wording B” presented the same images as Wording A, but instead asked which set of results the participant would choose. For iN6 Wording B, the means of the trials and the p-values were displayed. For iE6 Wording B, the means and the standard deviations of the trials were displayed.

The next three pairs of items (iN7/iE7, iN8/iE8, iN9/iE9) were paired for each suggesting a hypothetical different problem or issue to the participant that could have occurred in the contexts. Items iN7 and iE7 both proposed a type of confounding variable that hypothetically introduced bias to the data. Both items asked if the simulation still allowed for a valid answer to the research question. Items iN8 and iE8 suggested that the sample size for each study was too small and asked if running more trials of the respective simulation allowed for a small sample size. Items iN9 and iE9 both described a faulty measuring device, asking participants if the simulation still allowed for a valid answer to the research question.

Items iN10 and iE10 were paired due to both addressing how to conduct a replication study. For each item, the characters offered competing proposals, wherein one character argues to collect a new sample of data while the other erroneously argues to rerun the simulation.

Finally, data used for the first context were manually entered into Microsoft Excel and data for the second context were downloaded into Microsoft Excel. All computations and plots were created in the R statistical computing platform (R Core Team, 2019). Plots

were produced with the *ggplot2* (Wickham, 2016) and *gridExtra* (Baptiste, 2017) packages. Simulations were executed with the *mosaic* (Pruim et al., 2017) package.

3.3 Think-aloud interviews

Think-aloud interviews were executed to evaluate whether participants were interpreting the components of the instrument as intended, in order to update the instrument for the field test. The goal of think-aloud interviews is for participants to verbalize their thoughts as they process tasks and items (Gorin, 2006), which can inform the level of cognitive demand and any ambiguity of items (Haladyna & Rodriguez, 2013). The remainder of this subsection discusses the initial instrument versions used, interview participants, interview logistics, and the process to create the final blueprint.

3.3.1 Initial instrument versions for interviews

Four versions of the initial instrument were created based on the initial blueprint from Table 9, presented earlier. These four versions were considered the "first generation" of the instrument. For a complete draft of the initial instrument versions, see Appendix B1: Initial SUSIE versions for Think-aloud interviews.

To account for ordering effects of the two instrument sections, two of the four initial instrument versions presented the NHST section first and the other two versions presented the estimation section first. To evaluate the two different wordings of iN6 and iE6, two of the four instrument versions featured Wording A and the other two versions featured wording B. These two variables (order and wording) were crossed to create the four initial versions.

3.3.2 Participants and recruitment

The target participants were undergraduate or graduate students who had some experience with statistics and ideally had learned or worked with some type of statistical simulation. The objective was to interview students who either would be from the same courses that would be used for field test recruiting or at least had minimal statistical knowledge to provide effective feedback on the instrument. Students were recruited from mid-June 2020 through mid-July 2020, with recruitment overlapping with the execution of some interviews.

I recruited students from eight different course sections across three different courses at the University of Minnesota - Twin Cities and also colleagues. The three courses included two introductory simulation-based courses and one non-simulation-based course. One simulation-based course used a version of the Change Agents for the Teaching and Learning of Statistics (CATALST) curriculum (Garfield et al., 2012) and the other simulation-based course used the Lock 5 curriculum (Lock et al., 2013). Those two courses were the primary recruitment targets, as students learning with those curricula would also serve as the target participants for the field test. To recruit a sufficient number of participants, students were also recruited from a second semester non-simulation-based regression course. Typically, students in this course had taken a course with the Lock 5 curriculum as their first course. Also to recruit a sufficient number of participants, I recruited from among colleagues who were familiar with some type of statistical simulation.

All recruitment occurred via email. For recruitment from courses, I coordinated recruitment with consent and assistance from the respective course instructors. (I was the course instructor for three of the eight course sections.) Colleagues were recruited via

email. For the five course sections for which I was not the instructor, the instructors sent recruitment details via email on my behalf. To incentivize participation, a $10 Target gift-card was offered to one randomly selected participant who completed the interview. For interview recruitment materials, see Appendix E1: Instructor recruitment email template and Appendix E2: Participant recruitment email template. No consent form was needed given the exempt status assigned to this study by the IRB (STUDY00009911). Instead, a document was provided in recruitment emails with a summary of the study information and participant protections, included in Appendix E3: Information sheet.

3.3.3 Interview logistics

I executed all interviews over Zoom. All instrument versions were implemented in Qualtrics. For participants who agreed to be interviewed, a mutual one-hour block of time for the interview was established. To test different versions of the instrument, participants were randomly assigned to take a particular version of the instrument in advance.

An interview protocol was created to guide the interview process. In summary, each interview consisted of meeting the participant on Zoom; introducing the study and sending a link to the participant's randomly assigned version of the instrument on Qualtrics; participants reading all text on the instrument out loud and typing their responses to the items in Qualtrics; and providing a debrief about the instrument once sixty minutes had passed or the participant finished the instrument, whichever occurred first. Interviews were video and audio recorded on Zoom with the participant's consent. During interviews, participants were encouraged to continue to vocalize their thoughts throughout, to ensure they were sufficiently providing all feedback they had to offer. To avoid biasing

participants, corrective feedback to participant thinking during the interview was not provided. Any substantive questions the participant asked were reframed back to the participant. However, improvised follow-up questions were asked of participants at various points in some interviews, in order to glean deeper insight into some items. Additionally, I typed notes and observations about participant responses during each interview. For the complete interview protocol, see Appendix E4: Interview protocol template.

Several data sources were collected for each interview. This included the audio and video files from Zoom, the typed participant responses from Qualtrics, and the typed interviewer notes. Following interviews, I additionally created a typed transcript of each interview. No personally identifying information was collected other than the video and audio files themselves. All data were stored on a secured laptop and a secured shared Google Drive.

3.3.4 Iterative results and final blueprint

Twelve interviews were executed. One participant was from the simulation-based course with the Lock 5 curriculum, two participants were from the simulation-based course with the CATALST curriculum, five participants were from the non-simulation-based regression course, and four participants were colleagues. All interviews were executed without technical problems, except for one interview where only the audio was recorded due to the internet bandwidth being insufficient for allowing the participant's video to be activated. One participant shared personally sensitive information. The video, audio, and typed transcript files were edited to remove this personally sensitive information.

By the end of the twelve interviews, the initial versions of the instrument had been updated six times, leading to seven "generations" of the instrument having been tested in interviews. I created each new generation of the instrument following one or more interviews using the most recent generation. Some generations had more versions of the instrument than others. To create each generation, the most recent generations of the instrument were updated based on participant responses and interviewer notes up to that point. In summary, updates across generations included numerous item-wording changes, combining two items into one, changing details in the contexts, changing the layout of the instrument in Qualtrics, and adjusting the simulated outcomes used in the estimation context. For a reporting of results of the interviews and all consequent updates to the instrument, see Section 4.1 Think-aloud interviews. For the notes from each interview see Appendix C1: Complete interviewer notes. For detailed instrument updates that were considered or implemented across all generations and versions, see Appendix C2: Instrument changes throughout interviews. For the exact item wording tested in each generation of the instrument, see Appendix C3: Item history.

After applying final edits to the instrument based on feedback from interviews, the final blueprint was created, as shown in Table 10. The table lists the items from both instrument sections in the order that they appeared within each respective instrument section. This blueprint and version of the instrument were implemented in the field test, next discussed.

Table 10 *Final blueprint for SUSIE with item designation, text, and targeted facet(s) of real-world conflation*

NHST Section	Estimation Section
N1. Purpose (process, panacea) What is the purpose of randomization in the simulation (in terms of how it helps you answer the research question)?	**E1. Purpose** (process, panacea) What is the purpose of resampling in the bootstrap simulation (in terms of how it helps you answer the research question)?
N2. Center (product) After seeing the distribution of 500 trials, Researcher 1 offered this observation: *"Since the results center around 0, that suggests there is no difference in effect between fish oil and soybean oil in reality."* Do you agree with Researcher 1? Why or why not?	**E2. Calculate** (product) Your friends agreed to use 1.85 for the Standard Deviation of trials but disagreed on the Sample Estimate. Friend 1: *"The sample average of 6.35 minutes should be used."* Friend 2: *"The average of the 500 bootstrap trials, 6.10 minutes, should be used."* Which friend do you agree with? Why?
N3. Calculate (product) The researchers agreed to calculate a p-value but disagreed on the procedure. Researcher 1: *"Start at the average of the 500 trials, and then calculate the proportion of all trials larger than that."* Researcher 2: *"Start at the sample average difference of 7.7, and then calculate the proportion of all trials larger than that."* Which researcher do you agree with? Why?	**E3. Interpret** (product) To answer the research question, Friend 1 wanted to report the interval they calculated. However, Friend 2 said, *"You don't need an interval to answer the research question. You can report the average of the 500 bootstrap trials."* Do you agree with Friend 2? Why or why not?
N4. Interpret (product) Researcher 2 calculated a p-value of 0.014. However, Researcher 2 was unsure of how to answer the research question.	**E4. Center** (product) Suppose a third friend saw the plot of 500 trials and raised a concern. Friend 3: *"The average of 6.10 from the 500 bootstrap trials is different from the*

NHST Section	Estimation Section
Possible answer 1: *"The low p-value indicates that an average difference of 0 is not supported. There is evidence for a difference between the groups."* Possible answer 2: *"The low p-value indicates that the sample average difference of 7.7 from the data is not supported. There is no evidence for a difference between the groups."* Which answer do you agree with? Why?	*sample average of 6.35. When these two averages are different, this suggests the bootstrap average is more trustworthy than the original sample average."* Do you agree with Friend 3? Why or why not?"
N5. Trials (process, panacea) Consider the results from the smallest number of trials (100). Was there something gained by the other teams running more trials? If so, what did they gain and why? If nothing was gained, why not?	**E5. Trials** (process, panacea) Consider the results from the smallest number of trials (100). Was there something gained by the other groups running more trials? If so, what did they gain and why? If nothing was gained, why not?
N6. Replicate (process) Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results. Researcher 1: "We should randomly sample another 14 participants and put them through the same study that we just did." Researcher 2: "We can also run our simulation again with the data we already have, as that is as good as a replication of the study." Is the approach by Researcher 2 a valid way to do a replication study? Why or why not?	**E6. Replicate** (process) Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results. Friend 1: "We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample." Friend 2: "We can also run our bootstrap simulation again with the data we already have, as that is as good as a replication of the study." Is the approach by Friend 2 a valid way to do a replication study? Why or why not?
N7. Confound (panacea) Suppose that all of those in the fish oil group were taking a blood pressure	**E7. Confound** (panacea)

NHST Section	Estimation Section
medication, while all of those in the soybean oil group were not. Since each trial of the simulation randomly assigned the blood pressure values to the two groups, could the researchers still provide a valid answer to the research question? Why or why not?	Suppose that only morning departure times (in the range of 5am – 12pm) were included in the sample. Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?
N8. Sample Size (panacea) Suppose there was concern that the size of the original sample was too small. Did the fact that the researchers ran many trials of the simulation create a larger sample size for the study? Why or why not?	**E8. Sample Size** (panacea) Suppose there was concern that the size of the original sample was too small. Did the fact that your friends ran many trials of the bootstrap simulation create a larger sample size for the study? Why or why not?
N9. Device (panacea) Suppose that the medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have. Given the simulation that was run, could the researchers still provide a valid answer to the research question? Why or why not?	**E9. Device** (panacea) Suppose that the flight tracker at MSP airport was found to be faulty. It consistently recorded departure times as being later than it should have. Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

3.4 Field test administration

For the field test of SUSIE, participants were recruited beginning November 19th, 2020 and the data collection period occurred from December 1st, 2020 to December 17th, 2020. This time period was intentionally selected such that participating students would have been exposed to all intended content in their simulation-based courses, in order to have the minimum knowledge to complete the instrument. The remainder of this

subsection discusses recruitment of participants, field text logistics, and the development of the rubric to quantitatively score responses.

3.4.1 Participants and recruitment

The desired population was college students in introductory applied statistics courses with a simulation-based curriculum. This desired population was chosen due to the need to better understand student misconceptions arising from simulation-based introductory statistics courses. For practical purposes, the population of students from which students were sampled was limited to those who learned or are learning introductory statistics from either the CATALST (Garfield et al., 2012) or Lock 5 (Lock et al., 2013) curriculum in courses at the University of Minnesota - Twin Cities.

For the fall 2020 semester, there was one CATALST course comprised of four sections and one Lock 5 course comprised of four sections, including a total of 310 enrolled students across these eight sections as of December 1st, 2020. Both courses were offered by the Department of Educational Psychology (EPSY) at the University of Minnesota - Twin Cities. The CATALST curriculum was designed to cover basic ideas of statistical inference, using simulation and probability models. The course is typically taken by non-STEM undergraduate students who need a course that satisfies a mathematical thinking requirement. The Lock 5 curriculum covered introductory descriptive and inferential statistics, with the latter presented from both simulation-based and traditional theory-based perspectives. This course is typically taken by non-STEM graduate students who need a basic statistical methods course, although some undergraduate students are usually enrolled, as well.

To recruit participants, I worked with the seven instructors responsible for the eight course sections. Instructors were initially contacted to explain the study, request consent for recruitment, and explore the feasibility of offering extra credit to students as an incentive to participate. Following repeated correspondence, all seven instructors consented to assist with the study. Additionally, all instructors agreed to offer some form of extra credit to their students who participated, though the amount and type of extra credit was determined by and varied by instructor. For the initial recruitment email to instructors, see Appendix F1: Email template for recruiting instructors. For the follow-up correspondence with instructors, see Appendix F2: Email template for confirming study details with instructors.

Table 11 presents details about the courses included in participant recruitment. The two courses are identified by their curriculum name. (The University of Minnesota - Twin Cities course designations for the CATALST and Lock 5 courses are EPSY 3264 and EPSY 5261, respectively.) Instructors are designated by "C" or "L" to refer to the CATALST or Lock 5 curriculum, respectively, and a number for the course section taught. Two of the CATALST sections were taught by the same instructor and in the same format. This instructor treated these two sections as one large section, and thus sections 3 and 4 of CATALST were combined for the purposes of data collection and analysis. Two sections for each of the CATALST and Lock 5 courses were delivered in the remote format, while the remaining sections were delivered in the online format. For the remote format, class meetings occurred synchronously over Zoom. For the online format, all classwork and participation occurred asynchronously in the Canvas software platform. Student enrollment for each class was observed on December 1st, 2020.

Table 11 *Courses from which participants were selected for the field test*

Course	Section	Instructor	Class Format	Enrollment
CATALST	1	C1	Remote	44
CATALST	2	C2	Remote	45
CATALST	3 and 4	C3	Online	87
Lock 5	1	L1	Remote	19
Lock 5	2	L2	Remote	45
Lock 5	3	L3	Online	40
Lock 5	4	L4	Online	30

Note. Enrollment was observed as of December 1st, 2020.

Participant recruitment occurred from November 19th, 2020 to December 9th, 2020. Initially, students from the respective eight sections were contacted via email by their instructors on my behalf. Initial correspondence introduced the study to prepare students for the invitation to participate. On December 1st, 2020, an email invitation to participate in the study was sent to students by their instructors on my behalf. A reminder email to participate was also sent by instructors on my behalf on December 9th, 2020. Additionally, I made an in-class appearance over Zoom for each of the remote class sections, to introduce the study and take questions from potential participants. For the participant recruitment email and reminder email, see Appendix F3: Participant recruitment email template and Appendix F4: Reminder email template, respectively. For the in-class recruitment script, see Appendix F5: In-class additional recruitment script template. For the information sheet describing the study and participant protections that was sent to participants, see Appendix F6: Information sheet for field test.

3.4.2 Field test logistics

The field test was administered through Qualtrics. Access to the instrument was made available on December 1st, 2020 and closed on December 17th, 2020. Participants accessed the instrument via a link and password sent in recruitment emails. For one

section (EPSY 5261 Section 1), the instructor made the instrument an in-class activity, though students were not required to submit their responses if they did not want to participate in the study. For the remaining sections, participants completed the instrument on their own outside of class.

Upon clicking the emailed link and entering the password, participants were presented with a Qualtrics page describing the logistics of the study and how the extra credit would be awarded. Participants then selected their course section from a list on the second Qualtrics page. After these first two pages, participants were then randomly assigned one of two instrument versions. The first instrument version presented the NHST section first and the second instrument version presented the estimation section first. This random assignment was determined by Qualtrics, and a Qualtrics option was set that balanced out the sample sizes of the two conditions. Upon completion of the instrument, participants were redirected to a separate Qualtrics survey for the extra credit. While this survey asked for the participant's course section and email, these data were not linked with their instrument responses.

Several other Qualtrics options were set that impacted the participant experience with the instrument. Participants were allowed to click through pages without submitting responses. While this introduced the possibility that participants would submit blank responses just to easily earn extra credit, this was deemed the more ethical option, given the desire to avoid forcing participants to answer every item. Additionally, participants could not navigate back to previous pages. This was selected to prevent participants from changing their responses. To lower the stress on completion, participants could close the instrument once starting it and return to it later with their progress saved. However, if

open instrument attempts were left idle for more than one week, then the responses for that attempt were deleted. An option was also set to prevent the same participant from submitting more than one set of responses.

Data from each participant who a submitted a response to the instrument included their course section, their responses to the instrument, and other data internally tracked by Qualtrics such as the time duration. No personally identifying information was collected on the instrument. The version of the instrument that Qualtrics randomly assigned to each participant was automatically tracked as an internal variable in Qualtrics only visible upon viewing the participant data. For the extra credit survey, participant emails and the course section were initially collected. A list of participant emails was then sent to each respective instructor, and the extra credit survey data was deleted.

3.4.3 Rubric development

A quantitative scoring rubric was developed over several phases. In the first phase, a set of criteria and an exemplary response were drafted for each item. The scoring objective was initially established that responses would be scored a 1, 0.5, or a 0. A 1 signified meeting all stated criteria. A 0.5 generally signified meeting all the criteria but also stating something incorrect, denoting a response with meaningfully correct and incorrect assertions. All other responses were to be scored as 0. Non-responses (i.e., responses left blank) would not be scored, as no response was provided to be scored. Criteria for scoring a 1 or 0.5 were established based on the intent of the item and actual responses to the items from the Think-aloud interviews.

After the field test data were collected, the second phase consisted of putting the initial rubric through a validation procedure. First, 15 participants were randomly

selected from the field test data. Second, for each item, a colleague and I each independently scored up to 15 of the responses using the initial rubric. Third, the colleague and I compared our scoring decisions for each response, reconciled scoring discrepancies, and then updated the criteria in the rubric based on each reconciliation discussion. As a result of this process the 0.5 score was eliminated, to ease the difficulty in scoring the field test responses. Thus, a 1 signified meeting all criteria without making any meaningfully incorrect assertations. This meant that responses with both correct and incorrect assertions relevant to the content of the item would be a scored a 0. The target content for some items was also clarified, as detailed in the rubric. For the full rubric resulting from this process, see Appendix D1: Scoring rubric draft from validation.

The third and fourth phases consisted of updating the rubric while quantitatively scoring responses. (Details about the quantitative scoring process are covered in the next subsection.) In the third phase, I applied the rubric from the second phase to score all responses for the field test data. For numerous responses across all items, the rubric provided insufficient guidance for scoring. This was due to observing new types of responses that were not observed in the rubric validation procedure. This was also due to observing responses that were similar to what was observed in the validation procedure but were borderline in terms of meeting certain criteria. To eliminate this obfuscation and to sufficiently cover all types of responses, I updated the existing criteria and added new criteria to most items. In the fourth phase, I updated the criteria again while conducting the second qualitative review of the field test data. The rubric resulting from this review served as the final rubric that reflected the scores used for the quantitative analysis. For the full final rubric, see Appendix D2: Scoring rubric final draft.

3.5 Analysis of field test data

Data for analysis for this study consisted of participant responses to the eighteen items on the instrument and participant variables. Participant variables included their course, instructor, and which version of the instrument they completed (NHST section first or estimation section first). Data analysis consisted of two main areas: quantitatively scoring responses and qualitatively summarizing the themes of the responses marked as a 0. The general process for completing the data analysis consisted of cleaning the data, multiple rounds of quantitively scoring and qualitatively classifying responses, running the quantitative models, computing reliability measures, and summarizing the main qualitative themes. Notably, for the data analysis, the data were reviewed three times. In the first round, responses were quantitatively scored and the first attempt at qualitatively assigning themes to incorrect responses occurred. In the second round, a second attempt to qualitatively assign themes to incorrect responses occurred, while minor updates were applied to quantitative scores. In the third round, a final attempt at qualitative classification occurred. Each of these components of the analysis phase is next discussed.

3.5.1 Data cleaning

Cleaning the data for both the quantitative and qualitative analyses consisted of processing the raw field test data downloaded from Qualtrics to create an updated data file. First, unnecessary rows displaying instrument and Qualtrics metadata were deleted. Second, a participant ID variable was added. This was executed by randomly assigning a number from 1 to 193 without replacement to each row of the data, as each row corresponded to one participant.

Third, to ensure participants selected their correct course section on the instrument and extra credit, the sample sizes for each course section were compared between the instrument data and the extra credit data. This comparison revealed discrepancies, suggesting some participants either selected different course sections between the instrument and the extra credit survey or did not submit an extra credit survey response at all. For one participant, it was determined that they incorrectly selected their course section when completing the instrument. I manually changed this participant's course section in the updated data file that would be used for analysis.

Fourth, unnecessary columns that were automatically generated by Qualtrics but that would not be used for the analysis were deleted. The resulting file was the base file used for all quantitative and qualitative analysis. New variables were later added as columns as needed throughout the analysis process, to assist in conducting each analysis. This included a quantitative score variable and a variety of qualitative variables.

No participants were deleted from the base analysis file, as each analysis was governed by separate participant inclusion criteria. Therefore, all participants were included from the raw data file and then selectively excluded as needed for the given analysis. Participant exclusions are discussed below where relevant. The resulting sample sizes given these varying exclusions are further presented throughout Chapter 4: Results.

3.5.2 Quantitatively scoring responses

Quantitative scoring occurred by applying the scoring rubric to responses to all items, while iteratively updating the scoring rubric. The main scoring occurred in the first round of data review, with minor updates to scores occurring in the second round. The

scores from the second round were then used for the quantitative models and reliability measures.

In the first round of review, I executed a simultaneous quantitative scoring and qualitive coding procedure. (Qualitative coding is detailed in the next subsection.) For each non-blank response, the scoring rubric was applied and a score of 1 or 0 was entered. For blank responses a missing-value entry was entered. There were many unclear responses or responses for which the rubric was insufficient to provide a determination of a score. After an appropriate score for a given unclear response was determined, the rubric was consequently updated to reflect the entered score. Throughout this scoring and rubric updating process, numerous previously-scored responses needed to be rescored, to account for the updated rubric. At the end of the first round of analysis, all responses were assigned a 1, 0, or a missing-value designation.

The second round of review focused on a second attempt at assigning qualitative themes. However, during the qualitative review I noticed some quantitative scores that did not match the final rubric. Therefore, these scores were updated to reflect the rubric. This completed quantitative scoring.

3.5.3 Qualitatively classifying responses

The qualitative classification occurred over all three rounds of reviewing the data. However, only the results from the third round were used to answer the qualitative research question, as discussed below. In summary, the first round consisted of assigning general content themes to incorrect responses, the second round consisted of identifying responses suggesting a real-world conflation of the hypothetical simulation, and the third

round consisted of identifying the typical themes from the clearest examples of real-world conflation.

The first round of qualitative analysis was not used to inform the other rounds nor to answer the qualitative research question. This first round occurred while quantitatively scoring responses. The objective was to assign content themes to responses that were marked as a 0. The intent of this was to categorize the reasons for why responses were considered incorrect to uncover patterns in incorrect responses. Procedurally, this was executed by assigning one or more themes to incorrect responses immediately after they were scored as a 0. These themes were inductively created and frequently updated and refined throughout the first round of analysis. Following completion of this first round, I determined that the inductive approach was insufficient to accurately categorize the variety of incorrect responses, and that the resulting themes were not helpful in answering the qualitative research question. Instead, the first round of review helped inform a more targeted classification procedure for the second round of review.

In the second round, the objective was to rereview the incorrect responses and assign themes based on whether the response used language suggesting a real-world conflation with some aspect of the hypothetical simulation. Given the immense variety of incorrect responses and lack of clarity of language in many responses, a broad classification process was applied. This process consisted of making two classification decisions for each incorrect response. The first decision was whether the response used any language suggesting a real-world conflation. This was my subjective judgment based on the exact words used by the participant, the apparent intent of the response, and the nature of the given item.

To help make the initial determination about conflation language, the second decision involved assigning one of the three facets of simulation (Process, Product, or Panacea) to each part of a given response that suggested a real-world conflation. Thus, to make the initial determination about whether any conflation was present, some part of the response needed to answer the question, "Does this language potentially refer to a real-world process, a real-world product, or a power to affect the real world?" Once a response was determined to potentially involve a conflation, the one facet that seemed most relevant to the given language was assigned. Given the unclear and complicated nature of responses, assigning only one facet to a phrase was difficult. Myriad responses appeared to refer to multiple facets. For simplicity, only the facet that seemed most relevant to the response was assigned, though numerous responses potentially invoked two or all three facets. In the cases where multiple instances of potential conflation appeared in the same response, the phrases or words for each instance were separated and then assigned their own facet. Thus, some responses had several instances of conflation assigned to them.

Once all incorrect responses were rereviewed, the phrases from each response marked as suggesting conflation were initially aggregated by facet, within each item. Once grouped by facet, I attempted to further aggregate phrases suggesting conflation based on the similarity of their content. However, the high variety and often unclear nature of the language used in responses limited the ability to effectively aggregate language. Additionally, the difficulty in assigning just one of the three facets to each instance of possible conflation limited the validity of aggregating by facet. Moreover, both clear and unclear examples of conflation were considered equally in classification,

which lowered the defensibility in arguing for the clear presence of conflation in participant responses. Consequently, a third and final round of qualitative classification was proposed, informed by the subjectivity and difficulty encountered in the second round. For analysis comments and notes made during the second round, see Appendix G: Comments for Second Round of Classification.

In the third round, the objective was to identify the clearest examples of real-world conflation for each item and then identify the most typical, overarching themes for each item. Given the complexity of responses and exploratory nature of the study, the goal was to identify only a few broad themes illustrated with examples, instead of identifying all possible themes and tabulating frequencies. To accomplish this, I rereviewed the responses that were both marked as a 0 and also contained language suggesting the presence of a conflation. Two decisions were made for each such response. The first decision was whether the response offered a clear example of a conflation or an unclear example. The second decision was to assign one high-level theme to that response.

Determining clarity was a fairly subjective exercise, given the complexity of responses and the novelty of the content of this study. Generally, a response was marked as clear when the phrasing used words explicitly referring to a real-world product, process, or effect that also seemed unlikely to have multiple interpretations or meanings. In contrast, responses marked as unclear used real-world language but did so in an unclear, incomplete, or inexact manner, leaving room for multiple interpretations or meanings.

All responses with language suggesting conflation were assigned to a high-level theme. Themes were created inductively and iteratively while determining response clarity. Generally, themes were established by answering the question, "What does this response argue for, overall?", or "What is the main topic of this response as it pertains to a real-world conflation?" To simplify results and reporting, as few themes as possible were established for each item. Once themes were established within a given item, attempts were made to reuse that theme for the items that followed to reduce the number of themes to consider. For tracking purposes, the participant ID for a given response was listed under the theme to which it was assigned. The list of themes for each item and the participant IDs listed for each theme comprised the information used to answer the qualitative research question, discussed below, and are presented in Appendix H: Notes for Third Round of Classification.

3.5.4 Analysis for Research Question 1

Quantitatively analyses were executed to produce results to answer Research Question 1. This included producing basic descriptive and inferential statistics, running several regression models, and computing reliability measures. The data analysis was executed in the R statistical computing platform (R Core Team, 2020), with parts of the analysis and plotting executed with the *AICcmodavg* (Mazerolle, 2019), *sm* (Bowman & Azzalini, 2018), *ggplot2* (Wickham, 2016) or *psych* (Revelle, 2019) packages.

Inclusion criteria were established for the participants to be included in the quantitative scoring analysis. To be included in the scoring analysis, a participant needed to provide a text entry for each of the eighteen items, even if the text entry was nonsensical. Thus, submitting a blank response to any item excluded the participant from

the scoring analysis. Details on the resulting sample size for this analysis are presented in Section 4.2.

The outcome variables considered for the scoring analysis included scores on individual items, the total instrument score per participant, instrument section scores per participant, and the difference in instrument section scores (*difference score*) per participant. The instrument total score was computed by summing the two instrument section scores. Each instrument section score was computed by summing the individual item scores within the respective section. The difference score was computed by subtracting a participant's estimation section score from their NHST section score. The other variables considered as predictors or conditioning variables were course type (CATALST or Lock 5), instructor, and instrument order (NHST first or estimation first).

Basic descriptive and inferential statistics were first produced. This included computing the complete sample size used for the scoring analysis and the sample sizes of non-blank responses for each item and instrument section. Sample sizes contributing to the complete sample size for the scoring analysis were also computed when conditioning by course type, instructor, or instrument order received. (See Section 4.2 for a presentation of these sample sizes.) Histograms of the total score, two instrument section scores, and the difference score were produced.

As a measure of item difficulty and performance on the instrument, the percentage of the complete sample responding correctly to each item was computed. The percentage correct for each item was also computed when conditioning by course type, instructor, and instrument order received. The percentage correct was then compared between parallel items across the two instrument sections. For each of the nine pairs of

parallel items, this included computing the difference in the percentages correct, the standard deviation for the difference, and the confidence interval for the difference. Additionally, the Pearson correlation was computed for the percentages correct between parallel items.

Prior to regression modeling, basic descriptive and inferential statistics were computed for the instrument total score, the instrument section scores, and the difference score. For the total score, instrument section scores, and the difference score, the means, and standard deviations were computed with the complete sample and when conditioning by course type, instructor, and instrument order received. For the total score and difference score, the 95% confidence intervals were computed for the complete sample and when conditioning by course type, instructor, and instrument order. Additionally, the Pearson correlation between a participant's NHST section score and estimation section score was computed for the complete sample.

Next, the variables of total score and the difference score were separately modeled using ordinary least-squares (OLS) regression. These outcome variables were modeled to evaluate the extent to which they meaningfully varied by course type, instructor, or instrument order received. To compare models predicting the same outcome variable, the criteria used were the corrected Akaike Information Criterion (*AICc*), the "model weight" based on the *AICc*, and the model R^2.

As summarized, in Burnham et al. (2011), the uncorrected *AIC* essentially computes the amount of information loss relative to a true, unknown model. Thus, the smaller the *AIC* value for a model, the smaller the information loss relative to other

models, for a given set of candidate models predicting the same outcome. As discussed in Snipes and Taylor (2014), Akaike (1973) computed the *AIC* as follows,

$$AIC = 2K - 2log\ (L)$$

where *K* is the number of parameters being estimated and *log (L)* is the log-likelihood for the model. (In OLS regression, the parameters to be estimated include the intercept, the slope coefficients, and the error term.) To adjust for a small sample size, Hurvich and Tsai (1989) proposed the corrected *AIC*:

$$AICc = AIC + \frac{2K(K+1)}{n-K-1}$$

where *n* is the sample size for the model. While the model with the lowest *AICc* value may be considered the best model (Snipes & Taylor, 2014), multiple models may provide similar levels of evidence for being the best model among a set of candidate models (Burnham et al., 2011). One way to quantify the degree of model evidence is to compute the information distance, the relative model likelihood, and finally the model weight. Formulae adapted from Burnham et al. (2011) and Snipes and Taylor (2014) are next shown. The information distance, or delta, between each candidate model and the model with the lowest *AICc* is

$$\Delta_i = AICc_i - AICc_{min}$$

where $AICc_i$ is the *AICc* value for the i^{th} model and $AICc_{min}$ is the lowest *AICc* across candidate models. Next, delta can be converted to a likelihood value, as shown:

$$Relative\ likelihood_i = e^{-\frac{1}{2}\Delta_i}$$

Finally, the relative likelihood can be used to compute the model weight for a given model, as follows:

$$Model\ Weight_i = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum_{m=1}^{M} e^{-\frac{1}{2}\Delta_m}}$$

where the denominator is the sum of the relative likelihoods across all candidate models, and the numerator is the relative likelihood for the i^{th} model. The model weights across a set of candidate models sum to 1. Thus, the higher the model weight, the higher the amount of the available evidence that is assigned to the given model, given the candidate set of models.

Prior to evaluating models with the complete sample, models predicting the total score or difference score were run just for the CATALST or Lock 5 samples, separately. This occurred because the instructor variable was completely nested inside of the course type variable. Hence, both course type and instructor variables could not serve as predictors in the same model when using the complete sample. First, the following models predicting total score were estimated and compared within the CATALST sample and then estimated and compared within the Lock 5 sample, as shown:

$$Tot\ Score_i = \beta_0 + \varepsilon_i$$

$$Tot\ Score_i = \beta_0 + \sum_{j}^{J-1} \beta_j Instructor_i + \varepsilon_i$$

where $Tot\ Score_i$ is the total score for the i^{th} participant and *Instructor* is one of each of the *J* instructors for the i^{th} participant (with one instructor as the reference group, hence *J-1* in the summation). Second, the following models predicting the difference score were estimated and compared, as shown:

$$Diff\ Score_i = \beta_0 + \varepsilon_i$$

$$Diff\ Score_i = \beta_0 + \sum_{j}^{J-1} \beta_j Instructor_i + \varepsilon_i$$

where *Diff Score*$_i$ is the difference score for the i^{th} participant.

Based on the results of model comparisons when separating the sample by course type, I determined that the instructor variable was not a meaningful predictor. The instructor variable was therefore excluded from models using the complete sample. (For a discussion of this decision, see Chapter 4: Results.)

Next, a set of candidate models predicting the total score were estimated and compared, as shown:

$$Tot\ Score_i = \beta_0 + \varepsilon_i$$

$$Tot\ Score_i = \beta_0 + \beta_1 Course\ Type_i + \varepsilon_i$$

$$Tot\ Score_i = \beta_0 + \beta_1 Order_i + \varepsilon_i$$

$$Tot\ Score_i = \beta_0 + \beta_1 Course\ Type_i + \beta_2 Order_i + \varepsilon_i$$

where *Course Type*$_i$ is the curriculum for the i^{th} participant (with CATALST as the reference group) and *Order*$_i$ is the instrument order received for the i^{th} participant (with estimation-first as the reference group). The same set candidate set of models were estimated and compared for predicting the difference score, as shown:

$$Diff\ Score_i = \beta_0 + \varepsilon_i$$

$$Diff\ Score_i = \beta_0 + \beta_1 Course\ Type_i + \varepsilon_i$$

$$Diff\ Score_i = \beta_0 + \beta_1 Order_i + \varepsilon_i$$

$$Diff\ Score_i = \beta_0 + \beta_1 Course\ Type_i + \beta_2 Order_i + \varepsilon_i$$

Lastly, reliability measures were computed for completeness. Two versions of McDonald's omega were computed for the set of items within each instrument section

and for the entire instrument. This included omega-hierarchical (ω_h), which measures the degree of general factor saturation in a set of items, and omega-total (ω_t), which measures the total common variance shared among a set of items (Revelle & Condon, 2019). McDonald's omega was chosen as it requires fewer assumptions about item functioning compared to Cronbach's alpha when evaluating internal consistency (Dunn et al., 2014), and omega considers the factor structure inherent in a set of items for the computation and interpretation of the resulting measure (Revelle & Condon, 2019). For this study, the omega values were found with maximum likelihood estimation using a two-step factor analysis, with three factors extracted in the first step and one factor extracted in the second step. For the computation procedure and formulae, see Revelle (2019) or Revelle and Condon (2019). The item discrimination was additionally computed for each item. This was done by computing the point-biserial correlation between the scores on a given item and the total instrument score. Thus, a higher item discrimination indicates a stronger association between performance on the item and performance on the overall instrument.

3.5.5 Analysis for Research Question 2

To address Research Question 2, the responses using language suggesting a real-world conflation were evaluated and summarized. This included computing four different types of prevalence on a per-item basis, based on the determination of whether the response contained a clear or unclear example of a real-world conflation. These four computations were (1) the percentage of incorrect responses that were either clear or unclear examples of conflation, (2) the percentage of incorrect responses that were clear examples, (3) the percentage of all responses that were either clear or unclear examples,

and (4) the percentage of all responses that were clear examples. Furthermore, these percentages were subtracted between each of the nine pairs of parallel items. This was done to evaluate the difference in prevalence of conflation language between the NHST and estimation sections.

Next, the primary themes from each item from the third round of classification were summarized. This was done on a per-item basis by identifying the themes within each item that had relatively longer lists of participants who had provided clear examples of conflation. For some items, there were one or two themes that covered most of the clear examples of conflation, and thus, only these themes were reported. For items with much less consistency in participant themes, more themes were presented. To illustrate the complex nature of all themes, both clear and unclear examples of conflation were provided for each theme through selected participant quotes. This analysis was an arguably subjective exercise. Thus, the emphasis was placed only on identifying the most apparent themes with clear examples, instead of reporting the extent of all possible themes or numerical summaries.

Chapter 4: Results

This chapter summarizes results from the pilot interviews and field test relevant to the two research questions. Think-aloud interview results include an overview of the interviewer observations, interviewee feedback, and changes to the instrument. Field test results include sample sizes, quantitative scoring outcomes, reliability measures, and a summary of the most common themes from a qualitative review of the item responses.

4.1 Think-aloud interviews

Results from Think-aloud interviews were generated by the data from the twelve interviews. Data included video and audio of participant feedback while taking the instrument, instrument responses, and interviewer observations. The initial draft of the instrument that was administered to the first interviewee is presented in Appendix B1: SUSIE. Including the first interview, a total of ten versions of the instrument were administered over the course of the twelve interviews. For a description of differences among versions, see Chapter 3: Methods. The final instrument used in the field test was created following iteratively updating the initial and subsequent versions of the initial draft of the instrument. The final version is presented in Appendix B2: Final version of SUSIE for field test.

A summary of interviewer notes, observations, and instrument changes is discussed in the subsections that follow. Notes and observations were documented by the interviewer during the interview. For complete interviewer notes, see Appendix C1: Complete interviewer notes. For detailed notes on all instrument updates during the interview phase, see Appendix C2: Instrument changes throughout interviews. For the

exact item wording for each generation of the instrument with my comments summarizing the history of each item, see Appendix C3: Item history.

4.1.1 Instrument first generation feedback and changes

Two versions of the first generation were used in the first three interviews. Interviewees 01 and 02 used version B1, and Interviewee 03 used version C1. Several changes were made to the first generation following these three interviews.

Interviewee 01 only made it through the NHST section. Their responses indicated that the items were generally eliciting the intended content. However, they did not understand the logic of some of the prompts proposed by the characters in the instrument. The item about sample size elicited unrelated issues that appeared to distract from the intended content. For the items they completed, they provided thorough and accurate responses.

Interviewee 02 provided responses to all items. They generally provided inaccurate responses that did not match the intended content. They indicated they would not have been able to answer all the items without having taken a statistics class. In both the NHST and estimation sections they had difficulty differentiating between the item about the assumption behind the simulation and the item about how this assumption helps answer the research question. Additionally, in both sections, they took the context of some items into account for their responses to other items. This was not intended by the instrument blueprint. For example, they assumed the confounding variable discussed in one NHST item applied to all later NHST items.

Interviewee 03 completed all items. They missed the intended content on several items. Their responses informed multiple proposed changes. One issue was not

remembering the research question in the estimation section, as the research question was only displayed once on a separate page. Another issue was linking items that were not intended to be linked. In responding to the item about replication in the estimation section, they took the data problems from prior items in the estimation section into account. A third issue was giving authority to the teacher character in the estimation item about sample size, specifically because the character was a teacher. Giving authority to any character or their statements based on their title was not intended.

Changes to the first generation of SUSIE included Qualtrics reformatting, reordering some items, and rewording some items. Multiple interviewees dismissed the value of a replication study in the item about replication, due to assuming the original data were flawed as a result of the situations discussed in the items proposing hypothetical problems with the data. To reduce these cross-item linkages, the item about replication was moved from the tenth position to the seventh position in both the NHST and estimation sections. For each section, this moved the items about a confounding variable, sample size, and faulty device to the eighth, ninth, and tenth positions, respectively. To further reduce cross-item linkages and to reduce the number of words on each Qualtrics page, each of these three items was given their own individual Qualtrics page and the item wording was simplified. For the same three items, the "Supervisor" and "Teacher" characters were given unintended authority by Interviewee 03. These characters were thus removed to reduce the role of authority in considering their statements. Regarding the two items about sample size, all interviewees seemed unnecessarily fixated on the idea of a "small" sample size. This seemed to distract interviewees from the target content. The two sample size items were rewritten to ask

about trial size versus sample size more explicitly. Interviewee 03 could not remember the research question when responding to items later in the instrument. To help participants refer to the research question, the relevant research question was added to the top of each Qualtrics page. Lastly, the wording of the estimation item about replication was adjusted to be more parallel with the NHST item about replication.

4.1.2 Instrument second generation feedback and changes

A total of three different versions of the second generation were used across four interviews. Interviewees 04 and 07 took version B2, Interviewee 05 took version D2, and Interviewee 06 took version C2. Major instrument changes were applied following these four interviews.

Interviewee 04 completed eleven of twenty items on their instrument version. Their responses suggested that the items elicited the intended content. However, they broke the intended boundaries of two items. For the NHST item about trial size, they proposed selecting all four sets of results instead of selecting one set of results. For the NHST item about a faulty device, they proposed an accurate solution to overcoming the device error instead of critiquing the confounded data. Additionally, similar to other interviewees, they struggled to differentiate between the first two NHST items.

Interviewee 05 responded to fourteen of the twenty items on their version of the instrument. On many items, they discussed issues irrelevant to the items. They also struggled on several items. As with other participants the first two items in both sections were difficult for them. For the estimation item comparing the mean of the sample vs. simulated distribution, they questioned whether the magnitude of the difference was meaningful. Other responses prompted instrument changes.

Interviewee 06 only completed the estimation section. Their responses were typically addressed on the desired content of each item. They commented on a variety of aspects of the instrument. Their comments included assuming that the graph may have been needed to be memorized, adding units to the numerical values discussed in the estimation section, requesting that the research question relevant to the given section be presented like a reminder, and suggesting to increase the discrepancy between the sample and simulated distribution means. As with some other participants, for the item about a faulty device, Interviewee 06 discussed quantifying the measurement error in the data instead of critiquing the data.

Interviewee 07 responded to all twenty items on their version of the instrument. They expressed several concerns and difficulties. Some of their thinking throughout the NHST section seemed influenced by not trusting in the validity of the original simulation. Moreover, they used information from the NHST item about interpreting the p-value to consider changing their response to an earlier item about computing the p-value. This was not intended behavior for this instrument. Similar unintended item-linking occurred in the estimation section. For the NHST item about replication, they considered the sex of the study participants. For the wording of the estimation situation, they were concerned that the research question used the wording of "flight delays", when not all of the sample data contained delayed flights. For the estimation item about comparing the sample mean and simulated distribution mean, they commented that the prompt contained two different assertions loaded into the same prompt. This item was intended to only have one assertion to address. They also were concerned about the sample size difference between the NHST and estimation sections.

Changes to the second generation emphasized numerous item changes and some Qualtrics formatting changes. Across the first seven interviews, several participants struggled differentiating between the first and second items within each section. Thus, the first two items were replaced with a single item within each section that asked about the purpose of the given simulation.

For the estimation item comparing the sample mean to the simulated distribution mean, multiple interviewees thought the difference in means (6.35 vs. 6.39, respectively) was impractically small. Thus, a larger discrepancy was created, and the mean of the simulated distribution was changed to 6.10 minutes. To accomplish this, different bootstrap simulations were run until this larger discrepancy with the sample mean was obtained. Relevant plots and instrument text were updated to reflect this new bootstrap distribution mean. Additionally, one interviewee was concerned that the prompt asked about two issues in the same item: trusting the sample data and trusting the sample mean. Hence, the prompt was rewritten to only evaluate the sample mean.

Two interviewees disregarded the interval in the estimation item about reporting the interval vs. point-estimate to answer the research question for an unexpected reason. The interviewees argued that simply reporting the interval did not answer the research question. Hence, text was added to present both the interval and point-estimate as possible answers to the research question.

The text for the two items about replicating the study was adjusted to mitigate the concern about a low sample size. Focusing on a small sample size was not the intent of those two items. The phrase “with the same sample size” was added to clarify that the

sample size would not change for the replication study. Wording was also adjusted to ensure parallel phrasing.

Interviewee 07 discussed the sex of the study participants when answering the NHST item about replication, given concerns about generalizability of the blood pressure study results. To avoid distracting or influencing interviewees with specific demographic variables in the NHST blood pressure study, all demographic descriptors were removed. To further eliminate any concern about generalizability, participants for the blood pressure study were stated to have been randomly sampled from a generic population of interest. To ensure that both the NHST and estimation research questions referred to population, the phrase, "in the population", was added to the former.

Other minor text and design changes were implemented in Qualtrics to separate items and provide clearer instructions to participants. To prevent the item-linking that occurred with Interviewee 07, the NHST items about computing and interpreting the p-value were placed on separate Qualtrics pages. For the plots of sample data in both sections, some interviewees were not sure if that information needed to be memorized. A note was added stating that the graph did not need to be memorized, to avoid interviewees spending undue time and effort on memorization. The unit of "minutes" was added to the value of 6.35 throughout the estimation section, as Interviewee 06 wanted units added to the values. Interviewee 06 was also concerned that the research question being presented at the top of each page looked like a new question. Thus, the wording of the research question box on each page was adjusted to present it as a reminder of the original question. Lastly, "interval" was changed to "confidence or compatibility

interval" in the estimation section text presenting the interval formula, as some interviewees were not sure what a generic "interval" meant.

4.1.3 Instrument third generation feedback and changes

One version of the third generation was used in one interview. Interviewee 08 took version A3 and responded to all eighteen of their items. Their responses were only somewhat reflective of the intended content of the instrument. They questioned the accuracy of the randomization simulation in the NHST section. Generally, they seemed to lack familiarity with the randomization test, which may have interfered with responding to many items. For example, they referred to the randomization simulation as bootstrapping. On the NHST item about a faulty device, they went outside the intended bounds of the item by discussing the ethics of the researcher characters. Regarding the estimation item about answering the research question with an interval or point-estimate, they flagged the word "certain" in the argument for the point-estimate. This caused the interviewee to correctly choose the interval. While the point-estimate is the wrong response for this item, the word "certain" was not intended to influence participants. They additionally commented on the superfluous nature of some information provided on the instrument.

I discovered two minor unintended wording problems during the eighth interview. Since one item was removed for the third generation, the instrument and section introductions should have referred to nine questions instead of ten. This wording was changed to state that there were nine questions in each section. Additionally, Interviewee 08 did not explain their responses to the two items about trial vs. sample size. These items were unintentionally missing the follow-up prompt, "Why or why not?", which had

been in the first generation of these items. This prompt was readded. No additional changes were made, despite the feedback from Interviewee 08. I sought to pilot the third generation further before updating the instrument more.

4.1.4 Instrument fourth generation feedback and changes

One version of the fourth generation was used in one interview. Interviewee 09 took version D4 and completed ten of their eighteen items. They commented on several important aspects of the instrument. For example, they wanted to draft answers and then revisit them after seeing the other items. Similarly, they desired to consult a textbook. As with Interviewee 08, they flagged the word "certain" in the estimation item comparing an interval to a point-estimate. For the estimation item about replication, they did not endorse either statement presented to them. They argued that 500 trials were not enough trials in either proposal for a replication study. As with other interviewees, they questioned the accuracy of the NHST randomization simulation.

Changes to create the fifth generation primarily dealt with item wording in the estimation section. For the item about choosing between the interval vs. point-estimate for answering the research question, the word "certain" in the point-estimate appeared to be too strong a conclusion for multiple interviewees. To ensure the interval and point-estimate were presented with similar wording and strengths of veracity, the word "certain" was removed from the incorrect assertion about the point-estimate.

For the item comparing the difference in the sample mean vs. the bootstrap distribution mean, two new versions of the wording were created. The prior version compared the means with the idea of trusting one mean versus the other. Based on interviewee responses, proposing that the original sample mean could not be trusted did

not seem to elicit effective responses. For example, two interviewees stated that the sample mean could be trusted, but they each critiqued the sample mean in a way suggesting that they misunderstood the discrepancy between the sample and simulated means. Moreover, it was unclear if concepts other than trust would produce richer interviewee responses. For one new version, the idea of trust was replaced with the idea of accuracy. For the second new version, the item prompt proposed that the characters did something wrong since the sample and bootstrap distribution means were different.

For the two items about a replication study, one interviewee disagreed with both proposed approaches, which included the correct replication strategy. To account for disagreeing with both strategies, the prompts were adjusted to ask for the participant's preference of replication method. Additionally, the competing statements were rewritten for both items to use more parallel phrasing. The prior phrasing of the competing replication strategies used differing language between the statements, which may have distracted interviewees from the intent of the items.

Wording for both items equating trial and sample size was edited. Interviewee 09 fixated on the idea that bootstrapping was specifically intended to be implemented when there is a low sample size. Thus, the potential misconception of equating trial size with sample size was lost on this participant. To address this, the new wording for both items directly asked if running more trials created a larger sample size.

Lastly, when Interviewee 09 was presented with the computed interval in the item about reporting the interval vs. point-estimate, they went back to the previous item to verify that the computed interval was correct. To avoid interviewees wasting time on

unnecessary computations, a page break was added between the second and third estimation items in Qualtrics.

4.1.5 Instrument fifth generation feedback and changes

One version of the fifth generation was used in one interview. Interviewee 10 took version E5 and only completed the estimation section. This participant voiced unique concerns. This included anxiety about completing the instrument and coming up with sufficient responses. The interviewee looked up information at several instances during the interview. Generally, their responses discussed the intended content, including their incorrect responses.

Changes to the fifth generation involved significant item rewording. For the estimation item about answering the research question with an interval or point-estimate, Interviewee 10 was confused by the possibility of using the sample mean vs. simulated mean in computing the interval that was presented. This confusion unnecessarily distracted the interviewee from the intent of the item. To avoid future participant confusion, the prompt for the item was reduced to one statement about the point-estimate. Instead of participants choosing between the interval and point-estimate, the one statement incorrectly suggested to use the bootstrap distribution mean to answer the research question. This changed the focus of the item to evaluating whether using only the point-estimate was valid.

A new version of the two items about running more trials was created. Prior interviewee responses were mixed in quality and difficult to judge the accuracy of. The intended explanations to be elicited by the two items were not consistently appearing. Additionally, one interviewee broke the bounds of the items by choosing all four sets of

results, as a form of sensitivity analysis, as opposed to choosing one set of results. To reduce fixation on choosing a set of results, the two items were rewritten to ask about what was potentially gained by running more trials beyond the initial 100 trials. The intent of this rewrite was to directly elicit ideas about what benefits, if any, running additional trials offered.

For the two items about study replication, wording was adjusted in multiple places. One issue was having to indicate preference between competing replication strategies. Given the potential for disagreeing with both replication strategies, the prompt was adjusted to ask about agreement with only the incorrect statement, in contrast to choosing between the correct and incorrect statements. The order in which the correct and incorrect strategies appeared was made the same for both items. While the order of correct vs. incorrect competing statements was flipped between the NHST and estimation sections in other items, the content of the competing statements in the replication items was logical in only one order. Lastly, the prior version of the competing strategies both started with, “We should…”. I was concerned that emphasizing what should be done for a replication study might continue to cause problems for interviewees who disagreed with both presented strategies. The intent of the item was to evaluate whether interviewees considered rerunning the simulation as a form of a replication study. Thus, the incorrect statement began with, “We can…” instead of “We should…”. This changed the evaluation of the incorrect statement to be focused on whether rerunning the simulation was valid in any capacity, and not whether it should be the preferred strategy.

Finally, the research question reminder was added to the Qualtrics page with the estimation item about using the interval vs. point-estimate, as the reminder was mistakenly missing.

4.1.6 Instrument sixth generation feedback and changes

One version of the sixth generation was used in one interview. Interviewee 11 took version F6 and responded to all eighteen items. Overall, they responded correctly to most of the items. They provided several reactions and responses worth noting. When reviewing the bootstrap distribution of delay time means, they questioned whether the early departures were actually included in the data used to produce the simulated distribution. For their instrument version, a new phrasing of the NHST and estimation items about running more trials was used. This version of these items focused what on what was potentially gained by running more trials. This seemed to elicit a richer response than earlier phrasings of these items used in earlier interviews. As with other interviewees in the NHST section, they questioned what was done in the randomization simulation. Moreover, they desired to read the introductory text and simulation description again, after confronting the NHST item about increased trials later in the instrument.

Only a few wording changes were implemented to create the seventh generation. For the NHST item about computing the p-value, Interviewee 11 struggled with why only trials larger than a given value were being counted. This potentially was due to one-tail vs. two-tail confusion for the research question. Originally, the NHST research question was framed as a two-tailed question. To align the p-value computation in the item with

the research question, the NHST research question was rewritten to suggest a one-sided hypothesis test.

For the estimation item comparing the sample mean to the bootstrap distribution mean, Interviewee 11 suggested that the sample mean and the simulated mean could be quite different from the population mean, as the population mean is unknown. They also said the accuracy depended on the randomization. To discourage absolutist thinking on comparing these means, the word “accurate” was changed to “more likely to be accurate”.

Finally, for the estimation item about replication, Interviewee 11 incorrectly endorsed both strategies but preferred the correct strategy. To avoid endorsing both strategies, particularly for unclear reasons, the prompt was reworded to ask whether the incorrect statement was a valid way to do a replication study. The emphasis on validity was a more direct way to evaluate the incorrect strategy. The parallel change in wording was applied to the NHST item about replication.

4.1.7 Instrument seventh generation feedback and changes

One version of the seventh generation was used in one interview. Interviewee 12 took version G7 and verbally provided responses to all eighteen items but forgot to type responses to several items. They provided a mixture of correct and incorrect responses, occasionally not knowing how to explain their responses. They stated that they felt like they forgot much knowledge from their statistics classes. They additionally questioned whether a “Back” button was needed for the instrument. Having the research question for a given section at the top of each page helped alleviate this “Back” button concern. For the data presented in the NHST section, they spent extra effort in verifying what positive

and negative sample values meant. Regarding the NHST item about a flawed device, they explained a way to account for the measurement error, similar to other interviewees. For the two items about falsely linking trial size to sample size, they provided correct responses but could not explain their reasoning.

Only one item was adjusted from the seventh generation to create the final draft used in the field test. The estimation item comparing the sample mean and bootstrap distribution mean was reworded. Unfortunately, Interviewee 12 was unsure how to interpret the item and provided no clear response. Based on unsatisfactory responses to the versions of this item that compared the sample and simulated means with the idea of accuracy, I chose to use the idea of trust for comparing the means. However, prompting interviewees with, "we should not trust the sample average", elicited some undesirable responses in earlier interviews. Hence, the prompt was reworded to avoid strict evaluation of the sample mean and instead address whether the simulated mean was more trustworthy than the sample mean. Therefore, regardless of the degree of overall trust or accuracy a participant assumed about the sample mean or simulated mean, the item prompted them to directly compare their relative level of trust between the two means.

4.2 Field test

This section presents results from the analysis of the field test data. These analyses include the sample sizes used for each analysis, the outcomes from scoring responses based on the rubric, and the outcomes from classifying responses based on the presence of language suggesting a real-world conflation.

The initial data for analysis consisted of participant responses, instrument variables, and course variables for $N = 193$ participants. Table 12 shows the sample sizes

by item and the complete sample sizes for the two sections and overall instrument. At the item level, participants were included in the sample size if they provided any response, even if nonsensical. Of the original set of $N = 193$ participants, $n = 180$ participants or 93% of the original sample provided complete responses to the entire instrument. (For a participant to be considered a complete response for either section or the overall instrument, a text entry was needed for every item within the given section or both sections, respectively.) Item sample sizes varied from 189 to 193 participants. Two items (e02 and e04) included responses from the entire original sample.

Table 12 *Complete sample size by item and instrument section out of a possible N = 193 participants*

Grouping	Sample Size	Percentage of Initial Sample (%)
Item		
n01	190	98
n02	191	99
n03	192	99
n04	191	99
n05	191	99
n06	191	99
n07	190	98
n08	189	98
n09	190	98
e01	191	99
e02	193	100
e03	192	99
e04	193	100
e05	192	99
e06	190	98
e07	190	98
e08	192	99
e09	191	99
Completed Section		
NHST	184	95
Estimation	185	96
Completed Instrument	180	93

The scoring analysis and classification analysis used different sample sizes. The scoring analysis included the $n = 180$ participants who provided complete responses to the instrument. The complete sample size is discussed more in Section 4.2.1 below. The classification analysis was performed at the item level. Accordingly, the sample sizes for the classification analysis varied by item and correspond to the item-level sample sizes in Table 12.

4.2.1 Results from the scoring analysis

Results from the scoring analysis for the complete sample of $n = 180$ include descriptive statistics, inferential statistics, model testing, and three measures of reliability. Analysis outcomes focus on the total instrument score, instrument section scores, difference in instrument section scores (*difference score*), and item scores. The difference score was computed by subtracting a participant's estimation section score from their NHST section score.

The distribution of total scores, section scores, and difference scores are shown in Figure 2, Figure 3, and Figure 4, respectively. The maximum possible score on each section is nine, for a maximum instrument score of eighteen.

In Figure 2, the distribution of total scores is centered around nine points. Visually, the distribution is somewhat right skewed. Most scores are in the interval of five to fifteen points.

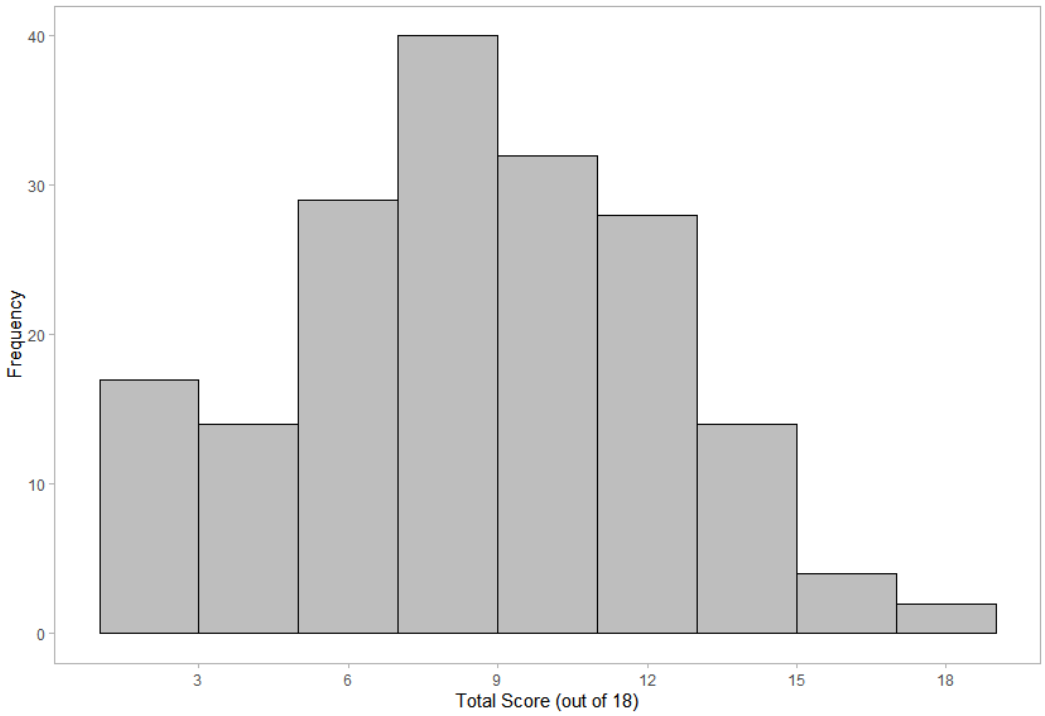


Figure 2. Histogram of instrument total scores

In Figure 3, the distributions of the two section scores are highly similar to each other. Both are centered around 4.5 points and are fairly symmetric. For each section, most scores are in the interval of three to seven points. Visually, there is little evidence for a performance difference between the two sections for the complete sample.

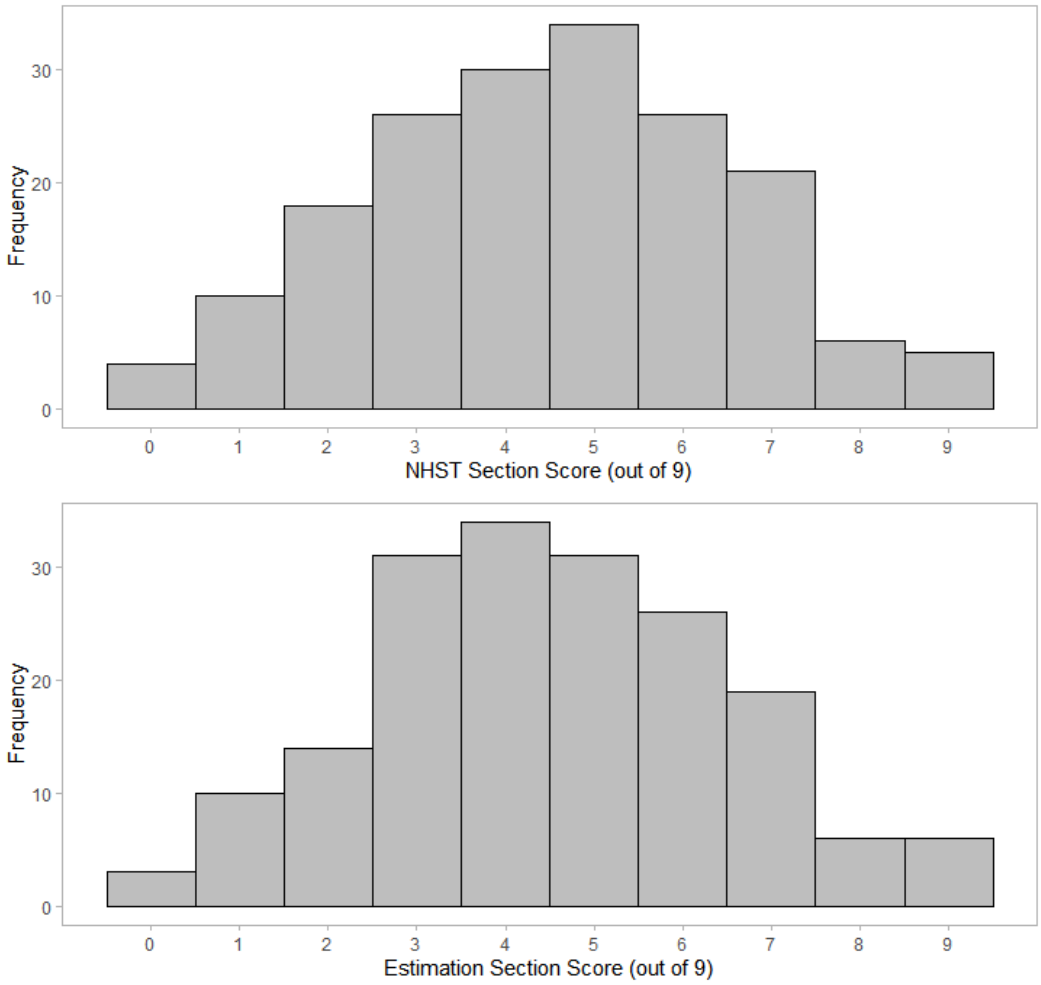


Figure 3. Histograms of total scores by instrument section

In Figure 4, the distribution of the difference between the scores on the two instrument sections is centered around zero and symmetric. Most difference scores are in the interval of negative two to positive two points. The maximum magnitude of the difference score is five. Of the complete sample (n = 180), 18% scored the same on both sections, 41% scored better on the NHST section, and 41% scored better on the estimation section.

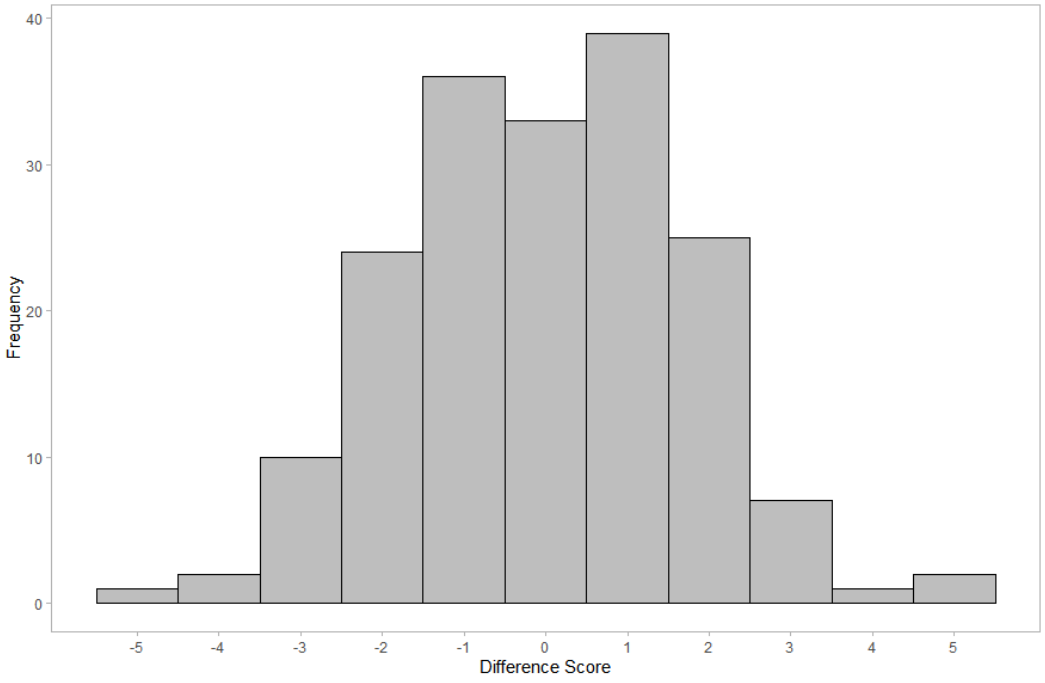


Figure 4. Histogram of the difference between NHST section and estimation section scores

The performance of the complete sample (n = 180) on each item is summarized in Table 13, Table 14, and Table 15 below. Performance was measured as the percentage of the sample that responded correctly to a given item. The percentage correct is presented and discussed for the complete sample overall and when conditioned by instrument order, course type, and instructor. Additionally, the performance difference between items that are parallel to each other between sections is presented and discussed.

Table 13 presents the percentage of the complete sample that correctly responded to each item, which is also conditioned by the order of the instrument sections. Item performance widely varied by item for the overall sample and when conditioning by instrument order. The two items that focused on the purpose of each simulation (n01 and e01) were the most difficult within each section, with only 11% and 24% of the complete

sample responding correctly, respectively. Thus, item n01 was most the difficult item on the entire instrument. The two items that focused on a flawed measurement device (n09 and e09) were the least difficult, with 82% and 78% of the complete sample responding correctly, respectively. Thus, item n09 was the least difficult item on the entire instrument. Of all eighteen items, at least 50% of the complete sample responded correctly to nine of them: both calculation items (n03 and e02), both interpretation items (n04 and e03), both items involving a confounding variable (n07 and e07), and the NHST item about trial vs. sample size (n08). Similar performance patterns are observed when conditioning by instrument order, with the two items focused on purpose (n01 and e01) again being the most or nearly the most difficult items in both conditions. Similarly, the two items focused on a flawed measurement device (n09 and e09) were the least or nearly the least difficult items in both conditions. Sample differences in the percentage correct for the same item between conditions were generally small. Only items e04, e05, e06, and e07 had differences over ten percentage points, favoring the NHST-first condition.

Table 13 *Item performance for the complete sample and by instrument order*

Item	Percentage Correct (%)		
	Complete Sample (n = 180)	Instrument Order	
		NHST First (n = 93)	Estimation First (n = 87)
NHST Section			
n01 (Purpose)	11	10	13
n02 (Center)	39	37	41
n03 (Calculate)	72	69	76
n04 (Interpret)	57	59	54
n05 (Trials)	30	27	33
n06 (Replicate)	42	46	38
n07 (Confound)	67	68	66
n08 (Sample Size)	51	52	49
n09 (Device)	82	82	82
Estimation Section			
e01 (Purpose)	24	22	26
e02 (Calculate)	52	54	49
e03 (Interpret)	68	67	69
e04 (Center)	28	33	23
e05 (Trials)	38	44	31
e06 (Replicate)	38	44	31
e07 (Confound)	77	82	71
e08 (Sample Size)	49	51	48
e09 (Device)	78	78	78

Table 14 shows the percentage correct by item, conditioned by course type and instructor. The pattern of item performance conditioned by course type was comparable to that of the complete sample and when conditioning by instrument order. Again, the two items about purpose (n01 and e01) were the most difficult within each instrument section for both course types, and the two items about a faulty device (n09 and e09) were the least difficult within each instrument section for both course types. Graduate students in a Lock 5 course performed better than undergraduates in a CATALST course on sixteen of the eighteen items. The largest difference was on the NHST item about

interpreting the answer to the research question using a p-value (n04), with the Lock 5 sample scoring 35 percentage points higher than the CATALST sample.

Table 14 *Item performance by course type and instructor for the complete sample (n = 180)*

Item	Percentage Correct (%)								
	Course Type		Instructor						
	CATALST (undergrad.)	Lock 5 (grad.)	C1	C2	C3	L1	L2	L3	L4
NHST Section									
n01 (Purpose)	8	14	0	7	15	12	10	16	19
n02 (Center)	26	53	34	15	28	56	52	50	56
n03 (Calculate)	71	74	55	78	77	81	81	75	56
n04 (Interpret)	40	75	45	30	44	81	95	59	75
n05 (Trials)	20	41	31	19	13	44	38	41	44
n06 (Replicate)	44	40	52	33	46	44	38	34	50
n07 (Confound)	58	76	59	52	62	81	67	81	75
n08 (Sample Size)	54	47	34	59	64	62	48	41	44
n09 (Device)	74	91	76	78	69	100	76	97	88
Estimation Section									
e01 (Purpose)	20	28	24	22	15	31	38	25	19
e02 (Calculate)	45	59	28	48	56	44	71	56	62
e03 (Interpret)	58	79	38	52	77	94	76	72	81
e04 (Center)	21	36	28	22	15	31	52	34	25
e05 (Trials)	27	49	28	30	26	56	62	31	62
e06 (Replicate)	34	42	34	30	36	56	43	34	44
e07 (Confound)	68	86	66	59	77	88	90	88	75
e08 (Sample Size)	45	54	41	52	44	62	62	41	62
e09 (Device)	71	87	59	85	69	88	81	94	81

Note. Sample sizes by course type: $n = 95$ (CATALST), $n = 85$ (Lock 5). Sample sizes by instructor: $n = 29$ (C1), $n = 27$ (C2), $n = 39$ (C3), $n = 16$ (L1), $n = 21$ (L2), $n = 32$ (L3), $n = 16$ (L4).

When conditioning by instructor, item performance again followed the trend of the overall sample and when conditioning by other variables. Considering each instrument section, both purpose items (n01 and e01) were either the most difficult or tied for the most difficult item across all instructors. Both items about a faulty device (n09 and e09) were either the least difficult or fairly less difficult than other items across all

instructors. Item performance was generally stable across instructors within the same course type, with exceptions. For example, the widest instructor gap on an item across CATALST instructors was 39 percentage points (78% vs. 39% for instructors C3 vs. C1, respectively) on the estimation item about interpreting results to answer the research question (e03). The widest gap across Lock 5 instructors was 31 percentage points (61% vs. 31% for instructor L2 or L4 vs. L3, respectively) on the estimation item about running more trials (e05). Across all instructors, the lowest performance was 0% on item n01 for instructor C1 and the highest performance was 100% on item n09 for instructor L1.

Table 15 summarizes the difference in performance between instrument sections for each set of parallel items. Items were paired between instrument sections by the focus of their simulation content. Based on 95% confidence intervals, evidence for performance differences in the larger population on a per-item basis is small to moderate. The intervals for five of the nine item pairs included a difference of zero percentage points. The strongest evidence for a difference was for the item pair about calculation. The interval was [13, 29] percentage points favoring the NHST section, with a point-estimate difference of 21 percentage points. Additionally, there was some inferential evidence for a performance difference in the item pairs about the purpose of the simulation (higher performance in estimation), the center of the simulated distribution (higher performance in NHST), interpreting the answer to the research question (higher performance in estimation), and addressing a confounding variable (higher performance in estimation). The point-estimates of these differences ranged in magnitude from 10 to 13 percentage points. Lastly, the Pearson correlation between the parallel item percentages was $r = 0.82$, further suggesting that parallel item differences were generally small.

Table 15 *Parallel item differences in the percentage correct for the complete sample (n = 180)*

Item Comparison	Difference in Percentage Correct (%) (SD)	95% CI
n01 – e01 (Purpose)	-13 (47)	[-20, -6]
n02 – e04 (Center)	11 (63)	[1, 20]
n03 – e02 (Calculate)	21 (55)	[13, 29]
n04 – e03 (Interpret)	-11 (62)	[-20, -2]
n05 – e05 (Trials)	-8 (58)	[-16, 1]
n06 – e06 (Replicate)	4 (49)	[-3, 12]
n07 – e07 (Confound)	-10 (56)	[-18, -2]
n08 – e08 (Sample Size)	1 (47)	[-6, 8]
n09 – e09 (Device)	3 (43)	[-3, 10]

Note. $r = 0.82$ for Pearson correlation between percentage correct on parallel NHST vs. estimation items

Table 16 presents the sample sizes, means, standard deviations, and 95% confidence intervals for the total instrument score out of a possible eighteen points. Sample sizes were fairly balanced within each conditioning variable. Sample sizes for CATALST and Lock 5 course types were reasonably balanced (*ns* = 95 and 85, respectively). The sample sizes of those who saw the NHST section first vs. those who saw the estimation section first were also fairly balanced (*ns* = 93 and 87, respectively). Higher sample size variation was observed across instructors, ranging from 16 to 32 participants. Instructor C3 had a relatively larger sample size due to teaching two course sections that were treated as one large section.

Instrument scores only slightly to somewhat varied when conditioning by course type, instructor, or instrument order. Across course type, graduate students in a Lock 5 course performed somewhat better than undergraduate students in a CATALST course. The sample mean score difference was 2.5 points and the confidence intervals of [9.6, 11.1] and [7.1, 8.6] for the Lock 5 and CATALST samples, respectively, do not overlap.

Given the difference between the CATALST and Lock 5 course types, differences between instructors of different course types were somewhat large. However, differences between instructors within the same course type were minimal. Within CATALST the range of the mean score by instructor was one point (7.3 vs. 8.3), and within Lock 5 the range was 1.4 points (9.7 vs. 11.1). Within course type, all confidence intervals between pairs of instructors overlapped with each other. Regarding instrument order, differences were also minimal. The difference between the two instrument orders was only 0.4 points, favoring those who saw the NHST section first, and the confidence intervals for each of the two instrument orders overlapped.

Table 16 *Summary statistics of the instrument overall and by course type, instructor, and order, out of a possible maximum score of 18 points*

Grouping	Sample Size	Mean Score (SD)	95% CI
Course Type			
CATALST (undergrad.)	95	7.8 (3.6)	[7.1, 8.6]
Lock 5 (grad.)	85	10.3 (3.4)	[9.6, 11.1]
Instructor			
C1	29	7.3 (4.0)	[5.8, 8.8]
C2	27	7.7 (3.3)	[6.4, 9.0]
C3	39	8.3 (3.4)	[7.2, 9.4]
L1	16	11.1 (4.3)	[8.8, 13.4]
L2	21	10.8 (2.8)	[9.5, 12.1]
L3	32	9.7 (3.2)	[8.5, 10.8]
L4	16	10.2 (3.5)	[8.3, 12.0]
Instrument Order			
NHST, Estimation	93	9.2 (3.6)	[8.5, 10.0]
Estimation, NHST	87	8.8 (3.8)	[8.0, 9.6]
Total	180	9.0 (3.7)	[8.5, 9.6]

Table 17 shows the means of the two instrument section scores and the means, standard deviations, and 95% confidence intervals for the difference score between instrument sections. Both the section means and the difference in section means are

conditioned by course type, instructor, and instrument order. Each instrument section had a maximum score of nine points.

Both instrument section means were typically around half of the total possible points. Differences in the mean instrument section scores were nonexistent to minimal. All 95% confidence intervals included a difference of zero points. Moreover, the Pearson correlation between NHST and estimation section scores was $r = 0.61$, suggesting moderate to high correspondence between the two instrument section scores. Thus, there is no clear evidence of mean performance differences between NHST and estimation for students outside this sample represented by these data.

Table 17 *Summary statistics of the section scores and the difference score by course type, instructor, and order*

Grouping	NHST Mean	Estimation Mean	Difference Mean (SD)	Difference 95% CI
Course Type				
CATALST (undergrad.)	3.9	3.9	0.1 (1.7)	[-0.3, 0.4]
Lock 5 (grad.)	5.1	5.2	-0.1 (1.8)	[-0.5, 0.3]
Instructor				
C1	3.9	3.4	0.4 (1.6)	[-0.2, 1.0]
C2	3.7	4.0	-0.3 (1.6)	[-0.9, 0.3]
C3	4.2	4.2	0.0 (1.8)	[-0.6, 0.6]
L1	5.6	5.5	0.1 (1.7)	[-0.8, 1.1]
L2	5.0	5.8	-0.7 (1.8)	[-1.5, 0.1]
L3	4.9	4.8	0.2 (1.9)	[-0.5, 0.9]
L4	5.1	5.1	-0.1 (1.6)	[-0.9, 0.8]
Instrument Order				
NHST, Estimation	4.5	4.7	-0.3 (1.9)	[-0.6, 0.1]
Estimation, NHST	4.5	4.3	0.2 (1.6)	[-0.1, 0.6]
Total	4.5	4.5	-0.0 (1.7)	[-0.3, 0.2]

Note. $r = 0.61$ for Pearson correlation between NHST score and estimation scores for the complete sample ($n = 180$)

Several ordinary-least-squares models were run that predicted the instrument total score, each of the two instrument section scores, or the score difference between the two

instrument sections. Results from these models are presented in Table 18 through Table 21 below. It should be noted that based on the sample means for the total score, section scores, and score difference between sections, there was minimal variation by course type, instructor, or instrument order. This suggests there may be limited utility in evaluating such models, but they are nonetheless included for completeness.

Prior to modeling the total instrument score or the difference in the mean instrument section score, the instructor variable was evaluated separately from course type. This is due to the instructor variable being completely nested inside of the course type variable. Thus, both the instructor and course type variables cannot serve as predictors in the same model.

Models predicting the difference score were run with the instructor variable for the CATALST and Lock 5 samples, separately. For comparison, an intercept-only baseline model was run. Results are shown in Table 18. In both course types, there was limited evidence to support differences in the difference score by instructor, given the lower *AICc* and much higher weight for the two baseline models. Thus, the instructor variable was not considered for predicting the difference score in the overall sample.

Table 18 *Model results predicting the difference score by instructor within each course type*

Model	K	AICc	Weight	R^2
CATALST Sample (n = 95)				
Baseline	2	371.81	0.71	-
Instructor	4	373.56	0.29	0.03
Lock 5 Sample (n = 85)				
Baseline	2	346.47	0.82	-
Instructor	5	349.54	0.18	0.04

Similar to predicting the difference score, the total score was first predicted separately by the instructor variable within each course type. Again, a baseline intercept-only model was used for comparison. Results are shown in Table 19. The lower *AICc* values and higher model weights strongly favor the baseline model in both course types. Again, the instructor variable was not considered for predicting the total score in the overall sample.

Table 19 *Model results predicting the total score by instructor within each course type*

Model	K	AICc	Weight	R^2
CATALST Sample (*n* = 95)				
Baseline	2	513.76	0.81	-
Instructor	4	516.62	0.19	0.02
Lock 5 Sample (*n* = 85)				
Baseline	2	450. 97	0.88	-
Instructor	5	455.01	0.12	0.03

Focusing on the complete sample, four models were run to separately predict the total score and the difference score. This included an intercept-only baseline model and models evaluating the variables of course type and instrument order. Table 20 shows results from predicting the difference score, and Table 21 shows results from predicting the total score.

Based on the results, model selection predicting the difference score is highly uncertain, with limited evidence beyond choosing the baseline model. The model with instrument order provided the lowest *AICc* and a model weight of 0.5. The baseline model and model with both course type and order also offered competitive weights of 0.22 and 0.18, respectively. Notably, the R^2 for the model with instrument order was only

0.02. This suggests limited evidence to prioritize the use of the instrument order variable above the baseline model. For completeness, the estimated equation is provided:

$$\widehat{DS}_i = 0.2 - 0.5 * NHSTfirst_i$$

where *DS* refers to the difference score and *NHSTfirst* = 1 for participants who saw the NHST section first and 0 for those who saw the estimation section first. Thus, for those who saw the NHST section first compared to those who saw the estimation section first, their difference score was a half-point lower (i.e., favoring better performance on the estimation section vs. NHST section), on average.

Table 20 *Model results predicting the difference score for the complete sample (n = 180)*

Model	K	AICc	Weight	R^2
Order	3	713.28	0.50	0.02
Baseline	2	714.94	0.22	-
Course Type + Order	4	715.30	0.18	0.02
Course Type	3	716.69	0.09	0.00

Based on the results for predicting total score, the model with just course type clearly has the most evidence. This is due to the lowest *AICc*, a model weight of 0.74, and a modest $R^2 = 0.11$. As reported in Table 17, the mean total scores for the CATALST and Lock 5 course types were 7.8 and 10.3, respectively, resulting in a sample mean difference of 2.5 points favoring graduate students in the Lock 5 course. These results match the coefficients in the estimated model predicting the total score by course type:

$$\widehat{TS}_i = 7.8 + 2.5 * Lock_i$$

where *TS* refers to the total score and *Lock* = 1 for participants in a Lock 5 course and 0 for participants in a CATALST course. The 95% confidence interval on the slope coefficient was [1.5, 3.5].

Table 21 *Model results predicting the total score for the complete sample (n = 180)*

Model	K	AICc	Weight	R^2
Course Type	3	962.85	0.74	0.11
Course Type + Order	4	964.94	0.26	0.11
Baseline	2	982.70	0.00	-
Order	3	984.14	0.00	0.00

Finally, reliability was evaluated with three measures. McDonald's ω_h and ω_t are discussed for the overall instrument and for the two instrument sections below. Item discrimination values are presented in Table 22.

There was some degree of general factor saturation (ω_h) and a fair amount of total common variance shared among the items (ω_t). For the overall instrument, $\omega_h = 0.38$ provided some evidence of a general underlying factor. Thus, 38% of the total variance in instrument scores was captured by a general factor. The eighteen items shared a fair amount of common variance, given $\omega_t = 0.80$. Thus, 80% of the total variation in instrument scores was "reliable" variation (Revelle & Condon, 2019). Within each section, there was also some evidence of general factor saturation and shared common variance. When comparing instrument sections, NHST items showed modestly higher general factor saturation than estimation items ($\omega_h = 0.47$ vs. 0.36, respectively). NHST items also showed a slightly higher amount of total common variance ($\omega_t = 0.71$ vs. 0.67, respectively). Accordingly, there is some evidence for measurement reliability, in terms of the instrument capturing variation in participants' scores that reflected actual differences in ability on understanding the content of the instrument.

Item discrimination values were fairly consistent. Within the NHST section, the lowest and highest item discrimination values were 0.38 and 0.60, respectively. Within

the estimation section, the lowest and highest item discrimination values were 0.33 and 0.52, respectively. Thus, item n04, which dealt with interpreting a p-value to answer the research question, provided the highest amount of discrimination. All eighteen items had discrimination values over 0.30. This suggests that all items contributed some degree of discrimination of participant performance to the overall instrument.

Table 22 *Item discrimination values*

Item	Point-biserial correlation with total score
NHST Section	
n01 (Purpose)	0.41
n02 (Center)	0.48
n03 (Calculate)	0.51
n04 (Interpret)	0.60
n05 (Trials)	0.40
n06 (Replicate)	0.46
n07 (Confound)	0.38
n08 (Sample Size)	0.45
n09 (Device)	0.39
Estimation Section	
e01 (Purpose)	0.35
e02 (Calculate)	0.50
e03 (Interpret)	0.43
e04 (Center)	0.43
e05 (Trials)	0.33
e06 (Replicate)	0.47
e07 (Confound)	0.41
e08 (Sample Size)	0.52
e09 (Device)	0.49

4.2.3 Results from the classification analysis

The classification analysis focused on identifying and summarizing responses that suggest a real-world conflation with the hypothetical simulation. The first set of results include a quantitative summary of the prevalence of conflation responses, covered in Table 23 and Table 24. The second set of results is a qualitative summary of the variety

of language used to suggest real-world conflation. Since only non-blank responses were included for classification, sample sizes for each item varied as shown in Table 12.

The prevalence of responses with language suggesting a real-world conflation was somewhat high across the entire instrument, with exceptions. The percentage of incorrect responses and the percentage of total responses that included language suggesting real-world conflation are shown in Table 23. These percentages are shown for all examples of conflation (which combined clear or unclear examples) and for only clear examples of conflation. When considering all examples, conflation responses were present in at least a quarter of total responses for nine of the eighteen items. For only clear examples, conflation responses were present in at least a quarter of total responses for only two of the eighteen items. Of incorrect responses, conflation language across all examples was present in the majority of responses for one item about purpose (n01), both items about linking sample size to trials (n08 and e08), one item about running more trials (e05), and one item about study replication (e06). Thus, for the remaining items, incorrect responses were marked as incorrect in a majority of responses due to reasons other than using conflation language. When considering responses with only clear examples of conflation, the two items linking trials to sample size (n08 and e08) had the highest prevalence among incorrect responses (44% and 45%, respectively). The lowest prevalence of responses using conflation language occurred in the NHST item about interpreting the results for the research question (n04), both for the collection all examples and for only clear examples.

Table 23 *Prevalence of responses suggesting real-world conflation*

Item	Prevalence (%)			
	Of Incorrect Item Responses		Of Total Item Responses	
	Clear or Unclear Language	Only Clear Language	Clear or Unclear Language	Only Clear Language
NHST Section				
n01 (Purpose)	62	30	55	27
n02 (Center)	34	26	21	16
n03 (Calculate)	13	5	4	2
n04 (Interpret)	1	0	1	0
n05 (Trials)	43	36	30	25
n06 (Replicate)	44	10	26	6
n07 (Confound)	20	15	7	5
n08 (Sample Size)	56	44	28	22
n09 (Device)	19	5	4	1
Estimation Section				
e01 (Purpose)	42	23	32	18
e02 (Calculate)	40	22	20	11
e03 (Interpret)	9	6	3	2
e04 (Center)	40	27	29	19
e05 (Trials)	52	33	33	21
e06 (Replicate)	63	14	39	8
e07 (Confound)	15	6	4	2
e08 (Sample Size)	61	45	31	23
e09 (Device)	11	9	3	2

Real-world conflation was generally higher in estimation items than NHST items. Table 24 presents the difference in the conflation prevalence between parallel NHST and estimation items, conditioned by clarity of language. Regardless of considering all examples versus only clear examples, conflation was more prevalent in estimation than NHST for five of nine item pairs (Center, Calculate, Interpret, Trials, Replicate, and Sample Size). However, the item pair about running more trials showed a higher prevalence of conflation for estimation only when considering all examples. Similarly, the item pair about a faulty device showed a higher prevalence of conflation for estimation only when considering clear examples. Regardless of considering all examples

versus only clear examples, two item pairs had a notably larger difference in prevalence compared to the rest: the pair about making a calculation (more conflation in estimation) and the pair about the purpose of the simulation (more conflation in NHST).

Table 24 *Difference in prevalence of responses that used language suggesting a real-world conflation between parallel NHST and estimation section items*

Item Comparison	Difference in Prevalence (%)			
	Of Incorrect Item Responses		Of Total Item Responses	
	Clear or Unclear Language	Only Clear Language	Clear or Unclear Language	Only Clear Language
n01 – e01 (Purpose)	20	7	23	9
n02 – e04 (Center)	-6	-1	-8	-3
n03 – e02 (Calculate)	-27	-16	-16	-9
n04 – e03 (Interpret)	-8	-6	-3	-2
n05 – e05 (Trials)	-9	3	-3	4
n06 – e06 (Replicate)	-19	-4	-13	-3
n07 – e07 (Confound)	5	9	3	4
n08 – e08 (Sample Size)	-5	-1	-4	-1
n09 – e09 (Device)	8	-3	1	-1

The remainder of this section summarizes the most common themes pertaining to examples of real-world conflation that were observed. Parallel items are introduced together, but the themes are primarily discussed on a per-item basis. Verbatim participant responses that illustrate clear and unclear examples of conflation relevant to a given theme are provided for each item. To preserve the original phrasing of each example response, spelling and grammatical errors were not corrected. Thus, "[*sic*]" was inserted adjacent to each language error below. For the complete analysis notes used to compile the summary below, see Appendix H: Notes for Third Round of Classification.

For the item about the purpose of the NHST simulation (n01), the themes with the most common occurrence of clear instances of real-world conflation dealt with internal or external validity. For example, this response targeted both forms of validity: “The use of randomization allows the simulation to have high external and internal validity. Having high external validity makes the results generalizable to the population, and having high internal validity allows researches [*sic*] to conclude causality.” [Participant 20]. In contrast, there were numerous responses that used language potentially suggesting a conflation but were unclear in this regard. These responses typically referred to “bias” or “accuracy”. One example was statements that referred to bias, as in this case: “The purpose of randomization is to reduce the possibility of bias in the experiment.” [Participant 163]. It is unclear what bias was being considered and whether the “experiment” was the simulation or the original study. Another example was statements that appeared to comment on the study design, either in lieu of the simulation or as if the simulation was part of the study design: “Randomization allows for there to be no bias, in that any preexisting conditions are picked at random.” [Participant 33]. The latter half of the response seems to describe the random assignment of the study but could potentially refer to the action of the simulation.

For the estimation item about purpose (e01), the most common types of statements were similar in theme to the NHST item about purpose, with exceptions. The clearest statements indicating a real-world conflation dealt with external validity. For instance, Participant 7 stated, “To best replicate the population and make the findings generalizable.” Other themes contained a preponderance of both clear and unclear statements. Similar to the NHST item, many responses emphasized removing problems

or improving something in the study situation, with varying degrees of clarity. Participant 185 offered a statement with a clear conflation: “The purpose of resampling [*sic*] the bootstrap simulation is to create a more accurate mean….” While it is unclear what “mean” the participant was referring to, “creating accuracy” is a direct statement about imbuing a real-world benefit through the simulation. In contrast, Participant 30 gave an unclear statement: “we want to ensure we have the “purest” samples in order to reduce the level of error”. It is unclear whether they referred to a real-world issue or simulation issue. While there was less of a presence of statements that seem to indicate a confusion with the study design compared to the NHST item, the estimation item instead elicited many responses about creating or increasing the sample size or data. A clear example was, “…resampling is an easier way to increase sample size rather than obtaining more raw data.” [Participant 167]. An unclear example was, “It allows you to create more samples” [Participant 126]. In the clear case, the participant directly referred to “sample size” in lieu of more data, while in the unclear case, the participant did not specify what “samples” refer to. Lastly, some statements with clear conflation indicated how bootstrapping allowed for testing or evaluating properties. These types of statements were not present in the parallel NHST item. For instance, “The purpose of resampling is to make sure your answer is accurate and your results weren't just a one time [*sic*] situation. Since the values are drawn randomly and then added back in you need to resample in order to prove validity.” [Participant 192].

For the two items about the center of the distribution, examples of clear conflation were easier to identify for the NHST item (n02) compared to the estimation item (e04). For the NHST item, simply agreeing with the prompt that the center of the simulated

distribution demonstrated no difference in the effect of the two oils suggested a conflation. The primary theme of most incorrect responses was assuming that a simulated distribution centered on 0 indicated no such effect. The degree of specificity in the explanation of these incorrect responses distinguished between clear and unclear conflations. A clear example of a conflation with sufficient explanation was, “I agree with Researcher 1 because since the distribution is centered at zero, it makes sense that there would be approximately no difference in the effects of fish oil and soybean oil.” [Participant 182]. Other misinterpretations of the center of the distribution and the results were less clear on the degree of conflation. For instance, Participant 122 stated, “I disagree with researcher [*sic*] 1. Although the data is centered around 0, the data is very spread out so you can't say there is no difference.” They referred to “data” centering around 0 and emphasized the spread. This may indicate a conflation, but more explanation is needed.

For the estimation item about the center, themes suggesting conflation were different than the NHST item. The most common type of example that clearly suggested a conflation involved explicitly affirming that the bootstrap average was more accurate or trustworthy than the sample average. Some of these responses included definitive reasons while others did not. As an unexplained example, Participant 12 stated, “yes, bootstrapping is more accurate than the sample average.” Some explanations added an additional conflation pertaining to increasing the sample size. To illustrate, Participant 36 indicated clear conflations involving sample size and accuracy: “I do agree, because the idea behind generating many bootstrap samples into a bootstrap distribution is to increase the overall sample size, and the large [*sic*] the sample size the closer our sample statistic

will be to the population parameter. The bootstrap interval will take into account the natural sampling variability and provide a better estimate of the actual population mean." However, most explanations involving more samples or data were unclear, in terms of the degree of conflation expressed. For instance, "Yes, because the bootstrap sample is an average of 500 trials while the sample average is just one trial.". [Participant 55]. The use of "trial" is ambiguous, suggesting that the actual sample is being treated with the same weight as a simulated trial. It is unclear what the participant viewed as being real or hypothetical.

The two items about calculation (n03 and e02) elicited fairly different amounts and types of responses suggesting conflation. Only seven responses to the NHST item potentially involved conflation, and only three responses clearly expressed a conflation. All three clear examples dealt with how the average from the trials provided the most accuracy. As Participant 68 argued, "I would agree with researcher [*sic*] 1 because the more trials you perform the more likely you are to get an accurate representation of the true mean." The remaining NHST responses with potential conflations were divergent in topic and difficult to understand.

For the estimation item about calculation, the typical responses with the clearest examples of conflation involved claiming benefits of using the bootstrap mean that the sample mean did not have. These benefits involved properties such as increased accuracy, among other properties. An example of this type of clear conflation was, "…The bootstrap test provides a more accurate estimate because it controls for the sample not be [*sic*] fully representative. It's closer to reality." [Participant 46]. Claiming that the bootstrap mean was tied to a larger sample size than the sample mean was a common

reason for this enhanced benefit. Referring to a larger sample size was also fairly prevalent. Using the phrase "sample size" suggested a clear conflation with actually increasing the sample size with the simulation, though it is unclear whether participants making this claim correctly understood the definition of "sample size". For instance, Participant 190 stated, "I agree with friend [*sic*] 2 that the 500 bootstrap sample [*sic*] should be used because it is more reliable because it is a larger sample size than the original trial."

The two items about interpreting the answer to the research question (n04 and e03) elicited few responses suggesting conflation. There were no clear examples of conflations in responses to the NHST item. The only response suggesting a conflation involved an ambiguous usage of the word "data". For the estimation item, there were only six responses with language suggesting a conflation. The clearest examples referred to the enhanced benefits of the bootstrap mean as the reason for reporting it instead of the confidence interval. However, not all examples clearly indicated that the participant understood the options in the item. Participant 63 chose the option to report the bootstrap mean instead of an interval, but their explanation did not reflect this. Moreover, conflations pertaining to increased sample size and enhanced benefits from the bootstrap simulation were present: "Yes. With a bigger sample you are able to give a more accurate interval of the average delay time."

Incorrect response suggesting conflation to the items about running additional simulation trials (n05 and e05) were difficult to understand and classify. A common type of response to either item was describing one or more benefits gained from an increase in sample size. These types of responses sometimes contained both clear and unclear

phrases within the same explanation. Example responses to each of these two items are presented next.

For the NHST item about running more trials, one example contained two clear conflations, where an increase in sample size was connected to a benefit suggesting external validity: “I think the teams running more trials got more accurate results because their sample included more people and in turn, they got results that were more representing [*sic*] of the real population” [Participant 151]. Other participants equated more trials to a smaller p-value or interval, due to the increase in sample size: “…As the trial size increases the average median also decreases. The p-value becomes smaller and closer to the center. The larger the sample size the smaller the p-value if the null hypothesis is false….” [Participant 68].

A variety of other benefits tied to an increase in sample size were stated, with varying degrees of clarity. For instance, Participant 89 provided a response potentially indicating one or more conflations but that is somewhat unclear, in terms of what the stated benefits actually pertain to: “Yes, something was gained. Larger sample sizes give more reliable results with greater precision”. Similarly, Participant 170 stated: “Using a larger sample size decreases the uncertainty of the results.” The usage of “uncertainty” and “results” is ambiguous. Other examples were difficult to parse, due to addressing several issues. The response from Participant 31 exemplified this: “The teams with higher trial numbers had more accurate data because a higher number of trials means better representation. The p- values *[*sic*]* would also be more accurate due to more data to trend with. Fewer trials do not have as much validity and could be biased.” Suggesting that fewer trials have less validity and more bias seems like a clear conflation about running

more trials. However, stating that having more trials equates to “more accurate data” less clearly indicates a conflation.

Example responses to the estimation item about running more trials similarly illustrated the degree of clarity in possible conflations. A clear example demonstrating multiple conflations discussed an increased sample size, a decrease in variability, and an increase in accuracy for estimating the mean: “The other groups that ran more trials compared to the first who only ran 100 gained an average that is closer to the actual, true average. They were able to accomplish this simply by taking larger sample sizes. Variability decreases as sample size increases, so a closer average to the actual one was able to be obtained in doing so.” [Participant 20].

In contrast, many participants discussed benefits that were used in unclear fashion, such as increased “accuracy”, “reliability”, or “precision”. For instance, Participant 8 ambiguously used the word “accuracy”: “larger sample size increases accuracy, larger number of bootstrap samples increases accuracy”. They moreover referred to both an increase in sample size and the number of bootstrap samples in the same response, though it is unclear whether they treated the two phrases as interchangeable. Another example highlights two unclear conflations: “yes, more data and more reliability” [Participant 61]. It is unclear whether they referred to purely hypothetical data or data tantamount to a real sample. It is additionally unclear what “reliability” pertains to. Another type of response linked an increased sample size to a generic decrease in uncertainty: “As the sample size increases, the uncertainty decreases. So by having more trials, more certainty is being gained.” [Participant 83] Without further explanation of what “uncertainty” the participant was thinking of, such as the

standard error, it is difficult to assume there was a conflation. However, some participants more explicitly discussed measures, such as the standard error or confidence interval, in ways that clearly suggested a conflation, as shown in this example: “Yes, by running more trials they were able to gather a more precise interval, aka [*sic*] making the range of the interval smaller. This is because when the sample size grows, it decreases the uncertainty in the data due to the fact that there is more data available which decreases the impact of outliers in a data set.” [Participant 176].

For the two items about replication (n06 and e06), identifying clear instances of conflation was challenging, due to the possibility that participants did not correctly understand what a replication study was in the first place. Clear instances of conflation were observed only with sufficient explanation related to some real-world aspect. For the NHST item, the clearest and most common examples suggesting conflation directly referred to the sufficiency of the simulation in place of a new study or similar real-world entity. Only six of these examples were found and each example was fairly different from the others. For instance, Participant 12 referred to new study participants, explicitly: “yes, if the data is randomized it would be the same as if we got 14 new participants”. Participant 105 similarly referred to new people but further tied their response to the role of new software: “Researcher 2 - new computer software allows researchers to resample without having to resample with a whole new group of people.” Discussing how real-world resources would be saved by rerunning the simulation also provided evidence of clear conflation, though there were only a few examples. Participant 111 illustrated both the idea of saving resources and how simulation was sufficient in lieu of new people: “It's always better to sample more people but in practice, time and money constraints will

limit this. 14 additional people will likely not make a huge difference…. Running a simulation is often the best we can do so it's valid."

Regarding unclear examples of conflation for the NHST item about replication, responses were highly variable in wording and topic. There were three typical themes that used unclear language potentially indicative of a conflation. One theme was simply agreeing with the premise that simulation is tantamount to replication, as shown by Participant 90: "Yes, I would say that is another way to replication [*sic*] study". The second theme described bootstrapping or sampling with replacement as a form of replication, as stated by Participant 15: "Yes, as long as a bootstrapping method is used. This will allow for multiple samples with replacement from a single random sample." Thus, there was obvious confusion about the simulation involved, as a randomization test (that does not use replacement) was described in the NHST instrument section. The third theme was invoking some aspect of randomness or randomization from the simulation as the justification for simulation as replication, as demonstrated in this example: "Yes, by randomizing the information already collected they will come up with additional data points." [Participant 191]. For these responses that contained limited explanation, described bootstrapping, or generically discussed something about randomness, it was not clear that there was a real-world conflation due to an apparent fundamental misunderstanding about the premise of replication or the simulation involved. In contrast, two participants who invoked bootstrapping did provide evidence of a clear conflation, as in the case of Participant 184: "No, unless they bootstrapped the simulation and ran it again, then they will have an accurate estimate of what could have occurred with a larger sample size, instead of using only 14 more participants." Their integration of the theme

discussed earlier of simulation being a sufficient replacement for new people added clarity to their response. Another notable theme was arguing that the simulation was valid for replication though not the preferred method. Participant 80 succinctly highlighted this: “It is valid, but it would be more effective to get new data.” As with most other responses, the extent of real-world conflation in this response is unclear, given the lack of explanation and potential misunderstanding about what replication is.

Comparable themes were observed for the responses to the estimation item about replication, though the variety of examples and language was much higher in contrast to the NHST item. Again, the clearest and most common examples of conflation involved referring to using the simulation in place of a real-world entity. Participant 168 provided a straightforward example of this: “Yes because in theory your bootstrap sample already contains enough variations of what could happen during a flight, so recreating it multiple times would essentially be the same as taking another sample from the population”. Many of these examples also integrated other themes, such as the necessity of having some random element or the behavior of bootstrapping. For instance, Participant 8 discussed the need for random sampling, the power of sampling with replacement, and the equivalence to a real sample: “if the first study was done by true random sampling, then that data should be representative of all departures, and sampling with replacement is representative of doing another actual sample. The only reason what [*sic*] it might be beneficial to do another study was if there was concern for confounding variables at the time of the original study”. As another example, Participant 63 argued that random selection negated the need to collect more data: “Yes. With being able to generate a larger sample by simulation you do not need to repeat collection of data to form another

original sample [*sic*] since the first collection was done by random selection it should be good." Moreover, their language suggests that running the simulation increases the sample size, by stating that the larger sample from the simulation also negates the need to collect more data. There were also several examples of clear conflation that invoked ideas of the simulation mimicking aspects of the real-world. For example, Participant 100 directly argued for the equivalence between the bootstrap and getting another sample by focusing on the replication of the process: "Friend 2's approach is a valid and accurate way to do a replication study because the bootstrap replicates Friend 1's process however many times we run the bootstrapping."

Regarding unclear examples suggesting conflation in response to the estimation item about replication, the themes and wording were wide-ranging and difficult to summarize. Given that the estimation section actually used bootstrapping in the simulation, there were naturally more examples arguing for the role of bootstrapping or sampling with replacement, in contrast to the NHST item. Most of these examples exhibited an unclear degree of conflation, however, due to the lack of explanation or uncertainty in whether participants understood replication. This was demonstrated by Participant 128: "Yes, because the bootstrap data uses with replacement [*sic*] so you can use the data that has already been replicated". Several unclear examples also integrated language that appeared to represent or potentially indicate mistaking the simulation for part of the study design, as shown in this response: "This is a valid way. The bootstrap simulation is also a random sample." [Participant 135]. Two themes with unclear examples of conflation were opposed to each other. One theme involved some aspect of the simulation staying the same. Referring to "steady data" and a "similar population" are

the entities remaining the same in this response: “Yes, since it is simulated towards a similar population with one area of departures, I believe that running another bootstrap simulation will be a good replication because they will only be using the steady data set they have from 2019.” [Participant 146]. In contrast, some argued that the replication is valid because some entity was different, as in this case: “Yes, because they will generate different results.” [Participant 178]. Finally, another typical theme for unclear examples of conflation referred to some aspect of randomness: “I think in this scenario friend [*sic*] 2 is correct. The sample was random and the sample size was greater than 30. The original data should be sufficient enough to use as replication.” [Participant 49]. Further notable in this example is arguing that a sufficiently large sample size ensured the validity of simulation as replication.

While there were fewer responses suggesting conflation compared to other items, for the two items about confounding variables (n07 and e07), a few types of clear conflation were observed. The nature of the prompt for these items aided in the clarity of identifying conflations. Merely agreeing with the item prompt and minimally explaining how the simulation overcomes a confounding variable to allow for a valid answer to the research question was sufficient to be labeled as a clear example. For the NHST item, most clear examples of conflation were tied to one of two themes. One theme involved explaining how the random assignment or experimental nature of the simulation allowed for a valid answer, despite a confounding variable, as stated by Participant 90: “Yes it would still be [*sic*] valid answer because the simulation randomly assigned the blood pressure values to the two groups.” Another example of this potentially indicated a misinterpretation of what the simulation provides, as the participant argued against

causality but argued for the validity of a correlation: “The introduction of the confound of additional blood pressure medication introduces potential doubts of treatment group as a causal factor in blood pressure outcomes, but any observe [*sic*] correlation still exists. Since each trial of the simulation randomly allocates these measures across the two treatment groups, the researchers can still provide a valid answer.” [Participant 64]. Another theme with clear examples of conflation simply argued that some aspect of randomness from the simulation ensured validity, as highlighted in this example: “Yes - the randomization process allows the researcher to estimate the variability within the study population and is used to simulate situations where not all variables can be controlled.” [Participant 71]. Unclear examples of conflation also referenced random assignment or randomness as ensuring validity but without directly addressing the simulation. This potentially indicated mixing up the simulation with the actual act of random assignment, instead of emphasizing the role of the simulation in ensuring validity. The generic response from Participant 164 illustrated this: “Yes, because of randomization they could provide a valid answer.”

For the estimation item about a confounding variable, only three responses provided clear examples of conflation. Two of these responses discussed variability or error measures from the simulation, as in this response: “It is possible. Bootstrapping uses sampling variation to account for things like - say, in this example - people being late for flights, heavy traffic to get to the airport, etc.” [Participant 162]. The other response argued that a valid answer could be provided, simply because the bootstrap simulation was random. Of the remaining responses suggesting conflation, most stated that a valid answer could be provided but did not provide intelligible explanations. One

response did clearly argue for validity from the shape of the bootstrap distribution, but they did not directly tie this validity to the simulation itself: “This is an outlier and it may make data less valid and more skewed, however since a bell shaped [*sic*] symmetric curve was still obtained in this study the 95% CI can be run and a valid answer can result.” [Participant 167].

For the two items equating trial size and sample size (n08 and e08), clear conflations were present; however, there was limited consistency in response themes. For both NHST and estimation items, there were three types of responses to the prompt, generally, which were difficult to separate. One type incorrectly agreed with the prompt that running more trials increased sample size. A second type correctly disagreed with the prompt but provided a reason suggesting conflation. A third type provided an explanation suggesting that they correctly disagreed with the prompt, even if they agreed with the prompt.

Most incorrect responses suggesting conflation for the NHST item were categorized under the first type above: incorrectly agreeing that running more trials increased sample size. Of these responses, there were numerous types of responses that indicated a clear conflation. Given that the prompt directly proposed a conflation, simply agreeing with the prompt and providing an explanation that used the words of the prompt or some interpretable explanation was sufficient to be considered a clear conflation. One typical type of response offered limited reasoning by simply restating the prompt in different words. Participant 38 provided an example of this with a clear conflation: “Yes. Running a [*sic*] many trials from a simulation tends to incase [*sic*] the number of cases as well as the sample size.” An unclear degree of conflation with limited reasoning is

illustrated by Participant 5: “yes because it was simulation”. For responses that gave detailed explanations, reasoning covered numerous themes that were difficult to categorize or define, such as various usages of the ideas of randomization, the population vs. the sample, representation, and other reasoning unique to single responses. Many responses additionally linked the trials to the increase in sample size directly, as explicitly stated in this example: “Yes because the trial size=the [*sic*] sample size”. [Participant 182]. Considering examples with more explanation, Participant 87 linked the sample size to the randomization of the results and the trials: “Yes it did create a larger sample size because it completely randomized the results while increasing the amount of trials substantially”. Taking a different angle, Participant 16 spoke to sampling without replacement and representation: “Yes because sampling without replacement allows one individual from the original sample to represent many individuals in the population.” In contrast, sampling with replacement and a slightly different usage of “representation” were stated by Participant 70: “Yes, it was resampling with replacement which is more representative of [*sic*] true population”.

For the NHST item, of responses that did not clearly agree with the prompt that running more trials increased the sample size, seven participants correctly disagreed with the prompt and six participants offered ambiguous explanations. For the former group of seven participants, all essentially gave an explanation that while the actual sample size was not increased, a larger sample size was simulated. Clear examples of conflation discussed a larger hypothetical sample size directly, while unclear examples of conflation did not. For example, Participant 4 offered a clear conflation, essentially arguing that the trials of the simulation provided insight into a larger sample size: “This does not create a

larger sample size, but it does allow for simulation of what results from a larger sample size would produce [*sic*].” On the other hand, Participant 53 only referred to a “replication”, which did not clearly indicate a conflation about insight into a larger simulated sample size: “it created more trials and therefore a larger replication but not a larger sample size”. Lastly, for the group of six participants with ambiguous disagreement with the prompt, clear conflation was observed in most cases, due to direct references to an increased simulated sample size. This is exemplified by the response from Participant 149 who, while agreeing with the prompt, only referred to a simulated larger sample size: “Yes, running many trials simulates a population and mimics a larger sample size.”.

For the estimation item about trials and sample size, most incorrect responses suggesting conflation explicitly agreed with the inaccurate prompt. Of these responses, a variety of explanations were given, similar to responses to the NHST item. Also, similar to the NHST item, simply agreeing with the prompt and providing minimally sufficient wording was enough to signify a clear conflation. One typical theme was providing limited, matter of fact reasoning. An example of this theme showing clear conflation was from Participant 41: “yes they had more data to work with.” When additional detail was provided, varying explanations were offered, instead of one or two definitive types of responses. These varying explanations involved reasoning about the trials, bootstrapping, different aspects of representation, assumptions about random sampling, and other reasoning unique to single participants. For instance, only one response that showed a clear conflation linked the population and sample in this manner: “Yes because we can assume the population is just many many [*sic*] copies of the sample.” [Participant 74].

Another example of clear conflation referred to the purpose of bootstrapping and an aspect of representation of the population: “This is what bootstrapping is commonly used for, to turn smaller data into data that represents the greater population. Therefore, I would say the answer [*sic*] correct.” [Participant 124]. In contrast, an unclear example of a conflation did directly address the changing sample size, despite explicitly agreeing with the prompt: “yes more trials mean [*sic*] more accuracy” [Participant 95].

For the estimation item, of responses that did not clearly agree with the prompt that running more trials increased the sample size, five participants correctly disagreed with the prompt and eleven participants offered ambiguous explanations. For the group of five participants, like the NHST item, the aspect of conflation in these responses involved referring to a larger simulated sample size. A clear conflation of this type was shown by Participant 127: “It didn't create a larger sample size, but it did simulate it.” For the eleven responses with ambiguous explanations, the clearest examples of conflation referred to a larger simulated sample size, without explicitly agreeing with the prompt. In this example, it appears the participant disagreed that running more trials increased the actual sample size without explicitly saying so: “Yes a larger sample size from simulation was simulated by bootstap [*sic*], as this is the goal of bootstrapping. The original sample size still is noted as the same however [*sic*] simulated bootstrapping number is noted in the reported results.” [Participant 167]. Nonetheless, they argued for a larger simulated sample size, indicating a clear conflation with what the power of what the simulation can demonstrate. Several unclear examples suggesting conflation were offered that did not clearly agree nor disagree with the prompt, as in the case of Participant 132: “Yes,

utilizing the bootstrap method allows us to have a smaller sample size, however the larger the sample size the more accurate our results will be."

Lastly, the two items about a faulty device (n09 and e09) elicited only a few clear examples of conflation. Only two responses illustrated clear examples for the NHST item. The most detailed response discussed the random assignment from the simulation but also appeared to hesitate in committing to the validity of answering the research question with a faulty device: "Yes, because the simulation utilized random assignment. However, they [*sic*] researchers would need to include a section discussing the limitations of this study, in terms of this extraneous variable influencing the results. They could say the results and highlight that the medical device gave higher readings for the soybean group than it should have." [Participant 32]. The other response was difficult to fully understand, which addressed some aspect of working beyond the observed data: "Yes because there were simulations that were run portraying more trials and results than just using the observed data, making it still valid.". [Participant 9]. For remaining examples suggesting conflation, all invoked some idea of randomness, random assignment, or random sampling, but did not explicitly relate this to the simulation. For instance, it is unclear if Participant 32 was thinking of the simulation or the random assignment at the outset of the study: "Yes, theoretically the inaccurate readings would have been evenly distributed through random assignment.".

For the estimation item about a faulty device, four of the five examples suggesting conflation were clear examples. Two of those four examples referred to the nature of bootstrapping as the reason for still being able to provide a valid answer to the research question, despite a faulty device. Participant 120 succinctly stated, "It is possible because

the purpose of bootstrap is to help [*sic*] those little error factors that could get in the way." In contrast, Participant 128 provided more detail: "yes, because the bootstrap data runs multiple simulations that run difference data that is similar to the observed data so they could still provide a valid answer that is relatively accurate.". In the other two clear examples, one referred to the "possible error" and "interval" from the simulation and the second referred to "running the sample again" to evaluate consistency, as explanations for still being able to answer the research question. The only unclear example of possible conflation did not directly connect to the simulation: "Yes because the data was made sure to be random" [Participant 160].

Chapter 5: Discussion

The Simulation Understanding in Statistical Inference and Estimation (SUSIE) instrument was developed and administered to better understand the nature and extent of how introductory statistics students conflate hypothetical simulations with the real world. Data were collected from students in courses that either used the CATALST curriculum or Lock 5 curriculum, as these are simulation-based curricula geared toward introductory students. The study consisted of developing the instrument, refining the instrument through Think-aloud interviews, administering the instrument to a sample of introductory statistics students, and conducting both quantitative and qualitative analyses.

The instrument was created over several phases. First, statistics students' misconceptions about simulations that were observed and discussed in the literature were synthesized. Misconceptions specific to conflating the hypothetical nature of simulations with the real world were then identified. Next, a framework describing three different facets of simulations where a conflation has been observed was proposed. A draft of the instrument and a blueprint were developed based on the framework and examples of conflation from the literature. The instrument presented two contexts that employed different methods for answering a statistical research question: (1) using a randomization test for null-hypothesis significance testing (NHST), and (2) using bootstrapping for estimation. Parallel items were written between the two instrument sections to explore how student performance and the nature of their misconceptions might vary between these two types of simulations.

The instrument was refined and piloted before being used for data collection. Informal discussions with statistics education advisors and colleagues led to feedback

used to finalize the first draft of the instrument. The instrument was then piloted in twelve Think-aloud interviews with undergraduate and graduate students. Numerous changes to the instrument were implemented, resulting in the final draft used for the field test. The field test consisted of collecting responses to the instrument from $N = 193$ introductory statistics students from one of eight different course sections across two course types at the University of Minnesota - Twin Cities.

Lastly, participant responses were scored, classified, and analyzed. A scoring rubric was developed and applied to assign scores to responses. Incorrect responses were classified over several rounds, with a primary emphasis on identifying clear examples of conflating simulations with the real world. Quantitative scores were analyzed with descriptive and inferential statistics, as well as a modeling process focused on the outcome variables of the instrument total score and the difference in scores between the two instrument sections. Qualitative themes were analyzed by identifying the primary themes arising from the clear examples of conflation apparent in the responses to each item. The remainder of this section answers the two research questions for the study, summarizes limitations, and discusses the implications for teaching and research.

5.1 Answering Research Question 1

The first research question was as follows: *To what extent are there quantitative differences in student understanding of the hypothetical nature of simulations when working with null-hypothesis significance testing vs. estimation?* To answer this question, select results from the quantitative analysis are summarized.

Based on modeling the difference in total scores between instrument sections, there was no clear evidence for a difference in performance between the NHST and

estimation instrument sections. Both the NHST and estimation section means were 4.5 points (out of a possible 9 points per section), resulting in a 95% confidence interval for the mean difference of [-0.3, 0.2]. The distribution of difference scores was symmetric and visually centered around 0 points, with most differences in the range of negative two to positive two points. Moreover, the Pearson correlation on the percentage correct between items that are parallel between the two instrument sections was $r = 0.82$. A similar lack in evidence for a difference was observed when conditioning by course type, instructor, and instrument order received. When predicting the difference score, only the predictor of instrument-order-received resulted in a model with an *AICc* weight larger than that of the baseline model (with no predictors), though the R^2 was only 0.02. Additionally, this effect was small, as those who saw the NHST section first scored only a half-point lower on the NHST vs. estimation sections, on average, with a 95% confidence interval of [-1.01, 0.01].

At the item level, there was some evidence for a difference in performance on parallel items between the NHST and estimation sections. Five of the nine item pairs had point-estimate differences in the percentage correct as large as 10 percentage points and 95% confidence intervals that excluded a difference of 0 percentage points. Three of these item pairs favored performance in the NHST section and two pairs favored the estimation section. Thus, any differences in performance at the item-level did not consistently favor NHST or estimation.

For completeness, a summary of the performance on the instrument as a whole is next summarized. The average score on the instrument was 9.0 points out of a possible total of 18 points (with a standard deviation of 3.7 points). Total instrument scores were

distributed symmetrically, with most scores in the interval of five to fifteen points. Importantly, since the overlap between the content of SUSIE and each course's learning objectives was not evaluated, the average performance of 50% on SUSIE may or may not indicate student difficulties with achieving the learning objectives for their given course.

Evidence from inferential statistics and the modeling process suggests that the total score only meaningfully varied by course type. For the CATALST vs. Lock 5 curricula, the mean total scores were 7.8 vs. 10.3 points, with non-overlapping 95% confidence intervals of [7.1, 8.6] vs. [9.6, 11.1], respectively. Furthermore, when modeling the total score, the highest model *AICc* weight was found when including only the course type variable. Based on the 95% confidence interval for the coefficient on course type in this model, this suggests that graduate students taking a Lock 5 course score 1.5 to 3.5 more points than undergraduates taking a CATALST course in a larger population of students matching the characteristics of this sample. This effect is potentially explained by the fact that enrollment in the CATALST course was mostly or all undergraduate students and the Lock 5 course was mostly or all graduate students. Within course type, differences in the total instrument score among instructors were minimal.

Performance at the item level varied widely, with the two items about the purpose of the NHST or estimation simulation the most difficult and the two items about a flawed measurement device the least difficult. This was generally true regardless of when conditioning by course type, instructor, or instrument order received. Looking across all items within the complete sample (n = 180), more than half responded correctly to the two items about calculating a result from the simulated distribution, the two items about

interpreting the results for the research question, the two items about addressing a confounding variable, the two items about a flawed device, and the NHST item about equating trial vs. sample size. Less than half of the complete sample correctly responded to the other nine items. Lower performance on an item may have resulted for one or more reasons. These include potentially high difficulty of an item, an item not aligning with course content, students not sufficiently learning course content relevant to an item, or students not sufficiently extending the course content to an item not covered in their course.

Finally, validity supporting such interpretation of the scores above is drawn from several sources. One main source is the instrument development process. The instrument was created based on a formal blueprint informed by a literature review of definitions of simulation and misconceptions held by students. Feedback from twelve Think-aloud interviews then led to seven rounds of edits to the instrument to ensure the items elicited responses as intended. Moreover, a scoring rubric was created, which was partially developed with a peer-validation process. Reliability measures are another source of validity evidence. Values of McDonald's omega indicate some degree of general factor saturation and shared variance among the items. Of the total variation in instrument scores, 80% was "reliable" variation (Revelle & Condon, 2019). Furthermore, item discrimination values indicate that all items separated higher and lower performers on the overall instrument to some degree, with all point-biserial correlations between items and the total score larger than 0.30.

5.2 Answering Research Question 2

There was evidence of participants conflating the hypothetical simulation with aspects of the real world in responses to nearly all items, with estimation items typically eliciting more examples of conflation than the parallel NHST items. The second research question was as follows: *What typical themes emerge that indicate students are conflating the hypothetical nature of the simulation with the real world?* To answer this question, select results from the qualitative analysis are summarized.

Based on the need for multiple rounds of qualitative review and the difficulty in classifying responses, an overarching theme of incorrect responses was the sheer complexity observed. The immense variety in response types and the unclear and inconsistent usage of certain words across responses limited the creation of a clear set of comprehensive themes that accurately categorized all response types. One issue was the apparent interchangeability of words such as "validity", "accuracy", and "reliability", among others, particularly across unclear responses. While participants may have had a specific definition in mind for such words, it was not always clear what that definition was from their responses. Moreover, terms such as "uncertainty", "data", and "random" were used in numerous ways, which in many cases, indicated potentially meaningful yet subtle distinctions in responses, despite similar wording. Such obfuscation in aggregating vs. distinguishing responses further limited the ability to create a clear set of classifications. An apparent lack of understanding terminology fundamental to interpreting the instrument items added further confusion in identifying conflations with the real world. Two main types of this were the idea of "replication" and the idea of "sample size". Replication was not clearly understood by some participants, which obfuscated whether they were exhibiting conflations when using language that suggested

they were. Similarly, “sample size” appeared to be used interchangeably with “samples” or “trials” in some instances, potentially indicating that a consistent interpretation of “sample size” was not shared by all participants.

Multiple explanations may be offered for why this variety and complexity of responses was observed. One explanation is that some students learning statistics with simulations simply hold a large variety of ideas and interpretations about the nature of statistical simulations, many of which are inaccurate or incomplete. Another explanation is that some students are still developing the appropriate definitions and usages of key statistical terms. Thus, they may hold a correct idea about a simulation but may not yet be able to clearly communicate it.

A third explanation is that expressing a conflation may require a minimum amount of knowledge about simulations. From a learning theory standpoint, this suggests that a baseline schema of understanding about simulations may be needed to have misconceptions about them. If this is the case, then if one does not possess a baseline schema, then one may not yet be able to exhibit misconceptions related to that schema. Consequently, encountering cognitive overload might result in explanations that are difficult to interpret, as was encountered. It was clear that at least some participants could not clearly express the ideas intended to be elicited by some items. This was demonstrated by Participant 167 who admitted to not understanding what a replication study entailed. Potential misuse of other expressions that seemed to indicate clear conflations might have been due to the partial understanding of terminology, such as “sample size”, as discussed above.

Many themes emerged when classifying incorrect responses at a macro level and focusing on clear examples of conflation. In one set of these themes, participants thought simulation afforded benefits that it actually does not, primarily invoking the Panacea facet of simulation. These benefits seemed to be rooted in real-world enhancements, including increased accuracy (of various measures), external validity, internal validity, or addressing confounding variables, among other benefits. Another set of themes discussed the outcomes from simulation as if they were equivalent to or actually a real-world entity. These primarily invoked the Product facet of simulation. The most notable theme of this type was that running the simulation can increase the sample size of a study. Another major theme related to the Product facet was that a randomization distribution centered on 0 indicated "no effect" in the fish oil context.

Another set of themes used language directly invoking real-world processes, which mainly reflected the Process facet. This was captured by the themes that simulation is tantamount to random assignment or a valid way to execute a replication study. Other major themes appeared related to multiple facets or were unclear in this regard, such as arguing that the center of the bootstrap distribution is more accurate or trustworthy than the actual sample mean.

Adding complication, some of these conflations were interconnected within the same response. For instance, many responses to the two items about sample size (n08 and e08) or the two items about running more trials (n05 and e05) argued for a variety of apparent real-world benefits from simulation as a result of the increased sample size that simulation supposedly creates. In contrast, numerous responses to the two items about the purpose of the simulation (n01 and e01) presented various real-world benefits from

simulation without explaining why. Thus, some types of misconceptions were observed repeatedly across items, but different items elicited different aspects of such misconceptions or led to some misconceptions being combined.

Most misconceptions and examples related to real-world conflation proposed in the literature were observed in participant responses. Regarding misconceptions pertaining to different ideas about replication, those examples observed by Rossman and Chance (2014) and Case and Jacobbe (2018) were clearly reflected in many responses. The former observed that students thought simulation replicates prior research to strengthen findings, which was a theme among incorrect responses to the two items about replication (n06 and e06) and somewhat present among incorrect responses to the two items about the purpose of simulation (n01 and e01). Additionally, Rossman and Chance (2014) argued that students did not understand that simulations were not generating new data. A primary theme from responses across several items was that both the randomization simulation and bootstrapping increased the sample size.

For Case and Jacobbe (2018), they observed student thinking that simulation can correct flaws in the original study. This was apparent in some incorrect responses to the two items about a confounding variable (n07 and e07), among others. Hodgson and Burke (2000) observed that when their students drew samples from a known population in a demonstration of the Central Limit Theorem, the students thought simulation was a "real-world strategy for finding the population parameter" (p. 94). While this study's participants were not interpreting a simulation with a known population in advance, thinking that simulation is a strategy to get a better estimate of the mean of a population was apparent. This was present in responses to the three items that involved the mean of

the bootstrap distribution (e02, e03, and e04) and the one item about the purpose of bootstrapping (e01). Participant explanations often cited the increased accuracy in the simulated mean or similar benefits from bootstrapping.

Previously observed misconceptions pertaining to the center of the simulated distribution were apparent in participant responses. Item n02 was created specifically based on the observation by Gould et al. (2010) that students misinterpreted the center of the simulated distribution based on a null hypothesis. Supporting this finding, the primary theme of incorrect responses to item n02 was that a simulated distribution from the randomization test that centered on a value of zero indicated no effect from the fish oil. As another type of misinterpretation of the center, students in Case and Jacobbe's (2018) study used the center of the simulated distribution as the starting point for computing the p-value. This misconception formed the basis of item n03. While performance was generally higher on this item compared to others (with 72% of the complete sample responding correctly), there were a few incorrect responses that matched this misconception. Clearer responses typically provided an explanation about how the center of the simulated distribution was more accurate or had some other beneficial property.

The example of the Discredit misconception as discussed by Case and Jacobbe (2018) was somewhat present. They found that at least one student disregarded the observed data as unlikely with a small p-value, instead of the null model. This idea formed the basis of item n04. While performance on this item was middling (with 57% of the complete sample responding correctly) and many responses did discredit the observed data explicitly, the explanations were often unclear. Thus, this misconception seems to have been present to a degree, but precisely determining this was difficult.

Moving beyond the literature, many novel misconceptions and examples of conflation were also observed. One major area of novelty was in the misconceptions when working with bootstrapping. In particular, a major theme was assigning enhanced accuracy or value to the mean of the bootstrap distribution over the mean of the original sample. In synthesizing the misconceptions from the literature, this was proposed to be the Discredit misconception from NHST applied to bootstrapping. Supporting this finding, numerous participants preferred and assumed a variety of benefits from the outcomes in bootstrapping. One noteworthy example of this was the explicit notion that a bootstrapping simulation is valid for a replication study but that a randomization simulation is not. Participants argued for other inaccurate ideas directly as a result of sampling with replacement. This suggests that there are real-world conflations and misconceptions unique to bootstrapping versus other types of simulations. Thus, some misconceptions about simulations may manifest differently in student thinking, depending on the simulation involved.

Another novel discovery was the array of benefits assigned to the randomization simulation, particularly in response to addressing the purpose of the simulation (n01). While Case and Jacobbe (2018) discussed how students may think that simulation can correct flaws in the original study such as a confounding variable, participant responses in this study covered a wider variety of specific benefits, such as increased internal validity, external validity, and reductions in study bias. In many responses, there additionally appeared to be confusion or conflation of the randomization simulation with the actual random assignment in the study design described in the NHST fish oil context.

Lastly, another novel set of findings pertained to responses to the items about running additional trials (n05 and e05). Across incorrect responses, several purported benefits of running more trials were stated, encompassing ideas such as increasing the sample size, lowering the p-value, reducing study bias, increasing validity, and providing more statistical power, among others. Given that performance on both items about trials was low (30% and 38% responding correctly to n05 and e05, respectively) and the variety of incorrect responses was high, this suggests that the conceptual understanding of simulated trials and their relation to real-world properties are messy.

5.3 Limitations

Several limitations in this study are next discussed. This includes the limitations to the generalizability of results, the role that curricula and software may have played in influencing results, the fuzzy nature of the conceptual framework used, and limitations from single-coder analyses.

Based on the curricula and conditions from which participants were sampled, results are generalizable to introductory statistics students in courses using the CATALST or Lock 5 simulation-based curricula at the University of Minnesota - Twin Cities. Additionally, these courses were administered remotely or online under the circumstances of the Covid-19 pandemic, where only synchronous-remote and asynchronous-online course formats were offered. In other semesters, these courses have typically been offered with synchronous-in-person and asynchronous-online sections. The factors of curricula, course delivery format, educational institution, and type of student might influence performance on the SUSIE instrument. Moreover, the sample might have been biased toward those who were motivated by the incentive to receive

extra credit in their course, potentially leading to an overrepresentation of lower or higher performers. Thus, if any of these factors are meaningfully related to instrument performance, then generalizability to students outside this population who do not match the participant characteristics in this sample that are meaningfully related to performance on the instrument is reduced. This is further explored next regarding the role of curricula and software.

The specific curricula and software used in this study may have played an influential role in the language used in responses. As has been explored elsewhere, the learning effects of using simulations may not be able to be disentangled from the content and pedagogy used in a learning environment (e.g., Hildreth et al., 2018; Tintle et al., 2011). This entanglement may be partly due to the intertwining of conceptual development with the technical instrument used. One perspective describing this intertwining is "instrumental genesis" (Artigue, 2002), where "[L]earning is seen as the simultaneous development of techniques for using artifacts, such as digital tools, and of domain-specific conceptual understanding, for example, statistical models and modeling" (van Dijke-Droogers et al., 2021, p. 236). Thus, the concepts that participants expressed in their responses may be inextricably tied to the simulation tools used in their respective course types. As van Dijke-Droogers et al. (2021) further stated, "[T]he specific intertwining of emerging digital techniques and conceptual understanding is unique for each digital tool and intended learning goal." (p. 258). For this study, students in the CATALST course sections used *TinkerPlots*^{TM} software (Konold & Miller, 2011) for simulations and students in the Lock 5 course sections used the *StatKey* online applet (Lock et al., 2017) for simulations. Between these two software programs, the

terminology that users learned, the layout of the simulation components on the screen, and other factors that users interacted with were somewhat varied. These variations might have influenced participants' conceptual development and language relevant to the content on the SUSIE instrument. Accordingly, generalizability of these results to students using other software may be limited. The role that the software played in participant responses may be highly meaningful but was not explored in this study.

Another study limitation was the interwoven nature of the facets of the conceptual framework and how that potentially interacted with participant responses. While the three facets of simulation (Process, Product, and Panacea) were used to create the blueprint, these facets seemed heavily intertwined or unclearly apparent in numerous participant responses. This obfuscated the classification of many incorrect responses, which reduced the capacity to evaluate responses against the blueprint or the role that the three facets played in participants' thinking. For instance, many incorrect responses arguably related to all three facets. Without further explanation by participants, it was challenging to discern if participants actually held misconceptions pertaining to all three facets. It is possible that some participants' misconceptions may have only stemmed from one facet, despite potentially sounding like they stemmed from two or three. However, through a combination of the nature of a given item and the response language used, it may be functionally impossible to reliably argue for the presence of only one facet in some responses. Moreover, the complexity of some of the target content may have required the consideration of multiple facets at once, reducing the utility of framing participant thinking or misconceptions with isolated facets. For instance, in conceptually dissecting whether a simulation is a valid replication of a study, it may be inaccurate to describe

misconceptions as only pertaining to a real-world product or a real-world process. Misconceptions about the products, processes, and panacea of benefits from simulations may be inextricably linked.

Finally, the use of a single scorer and coder reduced the extent of possible validity-evidence for outcomes from the quantitative scoring and qualitative classification. For reasons of practicality and limited research resources, I was the lone scorer and classifier. It is possible that scores and classifications may have shifted if they were subjected to multiple scorers or coders. Thus, scores and classification themes reported in this study may be subject to some degree of measurement error, adding some degree of additional uncertainty to both the quantitative and qualitative outcomes.

5.4 Implications for teaching

The results of this study suggest teachers who use simulation-based curricula should be aware that students may conflate aspects of the simulation with the real-world. These conflations may be multifaceted and difficult to identify. Ideally, there would exist effective steps teachers could take to first identify the presence of such misconceptions when students form them and then correct such misconceptions. Multiple theories and ideas proposed in the literature are helpful for framing potentially effective teaching behaviors, next discussed.

To consider how conflation-based misconceptions might emerge in the classroom, insight may be found by integrating the ideas of instrumental genesis, the unstable coordination of real-world vs. hypothetical perspectives, and guided discovery learning. First, this study evaluated students at the end of their introductory course. This suggests students may form their misconceptions throughout a course, which are then not fully

resolved before the course is completed. Second, given that statistical conceptual development appears intertwined with the software employed via instrumental genesis (van Dijke-Droogers et al., 2021), such misconceptions may be strongly tied to the simulation software of a course. These misconceptions may develop in tandem when students are introduced to each new function or idea that requires the use of the simulation software. Third, misconceptions pertaining to real-world conflation may be especially sensitive to learning with software. As Case and Jacobbe (2018) argued, their students' coordination of real and hypothetical perspectives was unstable, potentially due to the use of "representational media" that emphasized different aspects of the underlying systems (Johnson & Lesh, 2003; Lesh & Doerr, 2003). Fourth, as previously argued, learning with simulations can be a form of guided discovery learning (see Section 2.2.2). Particularly in an activity-based course, students may learn statistics with simulations through carrying out multiple steps with varying degrees of instructional guidance. For instance, both the CATALST and Lock 5 curricula in this study featured guided activities as a primary pedagogical activity. To increase the effectiveness of learning with guided discovery, sufficient instructional support is needed, which can include the use of guiding tasks or questions when interacting with simulations (de Jong, 1991; de Jong & van Joolingen, 1998). Hence, students may be repeatedly encountering moments of conceptual development via instrumental genesis for each guiding task, question, or step they encounter in the guided discovery process. These instrumental genesis moments might also provide opportunity for misconceptions to develop or be reinforced, contributing to inconsistent or unstable real-world vs. hypothetical perspectives.

To help alleviate conflation-based misconceptions, focusing on more formative assessment as students learn each new idea throughout a course with simulations may be beneficial. Previous calls to increase formative assessment to monitor student thinking and provide feedback to students have been made (Ben-Zvi et al., 2018; Garfield et al., 2011). Given a guided discovery learning environment and the repeated interactions with different parts of a simulation throughout an activity and course, it could be beneficial to add assessment checks to the steps of an activity or at the end of a key set of simulation tasks. Such learning checks could then be used in a formative manner to inform both the teacher and the student whether such misconceptions are held, allowing teachers to provide additional guidance before a misconception is further reinforced.

To aid in the development of formative assessment, content relevant to conflation-based misconceptions should be targeted in items comprising the formative learning checks. This could involve starting with the items or the target content from the items of SUSIE and adapting them to particular learning activities. However, given the complexity observed in participant responses to open-ended items and the significant effort and time needed to evaluate such responses, items in a selected-response or closed-ended format that are more efficient to score are presumably needed. The need for items that are efficient to score further connects to the implications for research, discussed next.

5.5 Implications for research

This study implemented an open-ended or constructed-response item format in the instrument. Scoring and classifying responses required non-trivial effort and time, as well as a rubric demanding detailed and intensive familiarity with the content. For more efficient future study of the target content and to consider efficient use of formative or

summative assessment in the classroom, creation of closed-ended or selected-response items is required. Additionally, different participants raised different ideas across incorrect responses; however, each participant was not evaluated on every incorrect idea that was stated by other participants. Thus, the prevalence of each misconception or type of conflation could not be reliably evaluated. For example, some participants stated that running more trials increased the statistical power. This conflation may have been held by more participants who simply did not include this reasoning in their response but would indicate such a misinterpretation were they to be asked about it. Thus, among other outcomes, this study served as an exercise in generating distractors for a selected-response version of the items in the instrument. This was exemplified by the two items about running more trials and the two items about the purpose of the simulation. Responses to these four items were wide-ranging, indicating a large set of responses that could be used as possible incorrect options, were these items to be converted to a multiple-choice format, for example. Developing a selected-response version of the instrument would allow for the target content to be covered in greater specificity with more items in a future study or formative assessment, given the knowledge of likely useful distractors from this study and the efficiency of scoring selected-response items. Moreover, such item development would contribute to filling the gap in modern statistics assessment, where items reflective of modern statistics curricula including simulation, are lacking (Tintle & VanderStoep, 2018; Ziegler, 2014).

To further support the development of selected-response items and to improve understanding about the interaction of simulations with real-world conflations, conceptual refinement is needed. The three facets of where a conflation in a simulation

may occur (Process, Product, Panacea) and previously observed misconceptions were used as a conceptual framework for developing the instrument. However, numerous responses were difficult to frame or clearly understand using this framework and these examples. This difficulty could be due to an inadequate framework or inadequate description of the target content. Drawing on the array and complexity of ideas elicited from the instrument, updating the framework that describes real vs. hypothetical interpretations of simulation may be crucial. Refining the target content and underlying framework should contribute to improvements in further instrument development and understanding the nature of these types of student misconceptions.

Another valuable area for research is connecting how certain misconceptions about simulations may impede the learning of other fundamental statistical ideas. For instance, incorrectly interpreting the center of the randomization distribution directly led to incorrectly answering the NHST research question (that the fish oil had no effect). While this example illustrates how a conflation may be conceptually adjacent to a statistical misunderstanding, some conflations may create problems further downstream, so to speak. For example, many participants prioritized the mean of the bootstrap distribution over the mean of the real sample. Hypothetically, a very large discrepancy between these means may indicate that the bootstrap simulation was executed incorrectly. On the one hand, this should lead a student to reevaluate the accuracy of the bootstrap simulation algorithm. On the other hand, a student who trusts the bootstrap mean more than the sample mean may not see this discrepancy as a problem and continue their analysis with the incorrect simulated outcomes. Continuing to unpack the role these types

of misconceptions about simulations play in students' understanding of other statistical ideas may help to diagnose various difficulties in how students learn statistics.

Lastly, a number of factors that could have influenced the nature of the responses to the instrument were not taken into consideration. These factors include the simulation type, curriculum version, level of student, course format, and type of pedagogy, among other factors. These types of confounding factors have been identified previously as playing potentially vital roles in the varying development and expression of students' understanding of statistics (e.g., Hildreth et al., 2018). Future studies should attempt to control for or describe how such factors influence the development of students' real vs. hypothetical perspectives of simulations. This understanding will be valuable for helping to identify the conditions under which certain misconceptions emerge and what steps may be taken to mitigate the negative impacts these misconceptions may pose.

References

Agresti, A., & Franklin, C. (2008). *Statistics: The Art and Science of Learning from Data* (1st ed.). Pearson.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281).

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1.

Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning*, *7*(3), 245–274.

Banks, C. M. (2009). What is modeling and simulation? In J. A. Sokolowski & C. M. Banks (Eds.), *Principles of modeling and simulation: A multidisciplinary approach* (pp. 3–24). John Wiley & Sons.

Baptiste, A. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics* (R package version 2.3) [Computer software]. https://CRAN.R-project.org/package=gridExtra

Beckman, M. D., Delmas, R. C., & Garfield, J. (2017). Cognitive Transfer Outcomes for a Simulation-based Introductory Statistics Curriculum. *Statistics Education Research Journal*, *16*(2), 419–440.

Ben-Zvi, D. (2000). Toward Understanding the Role of Technological Tools in Statistical Learning. *Mathematical Thinking and Learning*, *2*(1–2), 127–155. https://doi.org/10.1207/S15327833MTL0202_6

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of Statistics Learning Environments. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 473–502). Springer International Publishing.

Berry, K. J., Mielke Jr, P. W., & Mielke, H. W. (2002). The Fisher-Pitman permutation test: An attractive alternative to the F test. *Psychological Reports*, *90*(2), 495–502.

Biehler, R. (1997). Software for Learning and for Doing Statistics. *International Statistical Review*, *65*(2), 167–189. https://doi.org/10.1111/j.1751-5823.1997.tb00399.x

Bowman, A. W., & Azzalini, A. (2018). *R package "sm": Nonparametric smoothing methods* (2.2-5.6) [Computer software]. http://www.stats.gla.ac.uk/~adrian/sm/

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21–32.

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. J. (2013). Dynamic visualizations and the randomization test. *Technology Innovations in Statistics Education*, *7*(2).

Budgett, S., & Wild, C. J. (2014). Students' visual reasoning and the randomization test. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

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]. https://ufdcimages.uflib.ufl.edu/UF/E0/05/03/87/00001/CASE_C.pdf

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in learning inference from a simulation-based course. *Statistics Education Research Journal*, *17*(2), 9–29.

Centre for Education Statistics and Evaluation. (2017). *Cognitive load theory: Research that teachers really need to understand*. NSW Department of Education.

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education*, *1*(1). http://escholarship.org/uc/item/8sd2t4rr.pdf

Chance, B., & McGaughey, K. (2014). Impact of a simulation/randomization-based curriculum on student understanding of p-values and confidence intervals. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

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 (Ed.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS-10)*, Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on the Teaching of Statistics (ICOTS-7)*, Salvador, Bahia, Brazil. Voorburg, The Netherlands: International Statistical Institute. http://www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/7E1_CHAN.pdf

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.

Chen, B. (1999). *Transfer Between Different Contexts: Examining the Effect of Interaction* [Unpublished Master's thesis]. Rice University.

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

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1). http://escholarship.org/uc/item/6hb3k0nz.pdf

Dambolena, I. G. (1986). Using Simulation in Statistics Courses. *Collegiate Microcomputer*, *4*(4), 339–344.

de Jong, T. (1991). Learning and instruction with computer simulations. *Education & Computing*, *6*(3–4), 217–229.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, *38*(2), 105–134.
de Jong, T., Härtel, H., Swaak, J., & van Joolingen, W. (1996). Support for simulation-based learning: The effect of assignments in learning about transmission lines. In A. D. de Ilarazza & I. F. de Castro (Eds.), *Computer aided learning and instruction in science and engineering* (pp. 9–26). Springer.
de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179–201.
de Veaux, R. D., Velleman, P. F., & Bock, D. E. (2008). *Intro Stats* (3rd ed.). Pearson.
delMas, R., Garfield, J., & Chance, B. (1999). Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. *Journal of Statistics Education*, *7*(3).
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).
Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412.
Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26.
Engel, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
Erickson, T. (2006). Using simulation to learn about inference. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on the Teaching of Statistics (ICOTS-7)*, Salvador, Bahia, Brazil. Voorburg, The Netherlands: International Statistical Institute.
Fitch, A. M., & Regan, M. (2014). Accepting the challenge: Constructing a randomization pathway for inference into our traditional introductory course. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
Francis, G., Kokonis, S., & Lipson, K. (2007). Enhancing Student Understanding in Statistical Inference-Assessing the Effectiveness of a Computer Interaction. *IASE/ISI Satellite*.
Freund, J. E., & Williams, F. J. (1966). *Dictionary/outline of basic statistics*. Courier Corporation.
Galison, P. (1996). Computer simulations and the trading zone. *The Disunity of Science: Boundaries, Contexts, and Power*, 118–157.
Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment. *Teaching Statistics*, *31*(3), 72–77.
Garfield, J., delMas, R., Chance, B., & Ooms, A. (2006). *Assessment Resource Tools for Statistical Thinking*. https://apps3.cehd.umn.edu/artist/tests/index.html

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM - The International Journal on Mathematics Education*, *44*(7), 883–898.

Garfield, J., Zieffler, A., Kaplan, D., Cobb, G. W., Chance, B. L., & Holcomb, J. P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician*, *65*(1), 1–10.

Geyer, C. (2011). Introduction to markov chain monte carlo. *Handbook of Markov Chain Monte Carlo*, *20116022*, 45.

Good, P. I. (2006). *Resampling methods*. Springer. http://link.springer.com/content/pdf/10.1007/0-8176-4444-X.pdf

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.

Gould, R., Davis, G., Patel, R., & Esfandiari, M. (2010). Enhancing conceptual understanding with data driven labs. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.

Greca, I. M., Seoane, E., & Arriassecq, I. (2014). Epistemological issues concerning computer simulations in science and their implications for science education. *Science & Education*, *23*(4), 897–921.

Guidelines for Assessment and Instruction in Statistics Education College Report 2016. (2016). GAISE College Report ASA Revision Committee. http://www.amstat.org/education/gaise

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hesterberg, T. C. (2015). What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician*, *69*(4), 371–386. https://doi.org/10.1080/00031305.2015.1089789

Hesterberg, T. C. (1998). Simulation and bootstrapping for teaching statistics. *American Statistical Association Proceedings of the Section on Statistical Education*, 44–52. https://www.researchgate.net/profile/Tim_Hesterberg/publication/2714811_Simulation_and_Bootstrapping_for_Teaching_Statistics/links/54c6ff860cf238bb7d0a2ac0.pdf

Hildreth, L., 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).

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and. *Educational Psychologist*, *42*(2), 99–107.

Hodgson, T., & Burke, M. (2000). On simulation and the teaching of statistics. *Teaching Statistics*, *22*(3), 91–96.

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Johnson, T., & Lesh, R. A. (2003). A models and modeling perspective on technology-based representational media. In *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 265–278). Lawrence Erlbaum Associates.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Knapp, H. R., & FitzGerald, G. A. (1989). The antihypertensive effects of fish oil. *New England Journal of Medicine*, *320*(16), 1037–1043.

Konold, C., & Miller, C. (2011). *TinkerPlots* (Version 2) [Computer software]. Key Curriculum Press.

Lane, D. M., & Peres, S. C. (2006). Interactive simulations in the teaching of statistics: Promise and pitfalls. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on the Teaching of Statistics (ICOTS-7)*, Salvador, Bahia, Brazil. Voorburg, The Netherlands: International Statistical Institute. https://pdfs.semanticscholar.org/579a/c6326054fe4932a1cc979815763fce6bce1f.pdf

Lane, D. M., & Tang, Z. (2000). Effectiveness of simulation training on transfer of statistical concepts. *Journal of Educational Computing Research*, *22*(4), 383–396.

Lesh, R. A., & Doerr, H. M. (2003). Foundations of a models and modeling perspective on mathematics teaching, learning, and problem solving. In *Beyond constructivism: Models and modeling perspectives in mathematics problem-solving, learning, and teaching* (pp. 3–34). Lawrence Erlbaum Associates.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics (ICOTS-6)*, Cape Town, South Africa. Voorburg, The Netherlands: International Statistical Institute. https://iase-web.org/documents/papers/icots6/6c1_lips.pdf

Lipson, K., Francis, G., & Kokonis, S. (2006). Developing a computer interaction to enhance student understanding in statistical inference. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on the Teaching of Statistics (ICOTS-7)*, Salvador, Bahia, Brazil. Voorburg, The Netherlands: International Statistical Institute.

Lipson, K., Kokonis, S., & Francis, G. (2003). Investigation of students' experiences with a web-based computer simulation. *Proceedings of the 2003 IASE Satellite Conference on Statistics Education and the Internet*.

Lock, R. H., Lock, R. H., Morgan, K. L., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the power of data*. Wiley Hoboken, NJ. http://wileyplus.wiley.com/statistics-unlocking-the-power-of-data-2nd-edition/

Lock, R., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2017). *StatKey* (2.1.1) [Computer software]. https://www.lock5stat.com/StatKey/

Maurer, K., & Lock, D. (2016). Comparison of Learning Outcomes for Simulation-based and Traditional Inference Curricula in a Designed Educational Experiment. *Technology Innovations in Statistics Education*, *9*(1). http://escholarship.org/uc/item/0wm523b0.pdf

Maxara, C., & Biehler, R. (2007). Constructing stochastic simulations with a computer tool—Students' competencies and difficulties. *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education*, 762–771. http://www.erme.tu-dortmund.de/~erme/CERME5b/WG5.pdf#page=79

Mayer, R. E. (2002). Cognitive theory and the design of multimedia instruction: An example of the two-way street between cognition and instruction. *New Directions for Teaching and Learning*, *2002*(89), 55–71.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14.

Mazerolle, M. J. (2019). *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)* (2.2-2) [Computer software]. https://cran.r-project.org/package=AICcmodavg

Mendoza, S., & Roy, S. (2018). Assessing retention of statistical concepts after completing a post-secondary introductory statistics course. In M.A. Sorto (Ed.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS-10)*, Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, *61*(1), 1–15.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, *10*(1), 1–20.

Mills, J. D. (2004). Learning Abstract Statistics Concepts Using Simulation. *Educational Research Quarterly*, *28*(4), 18–33.

Mooney, C. Z. (1997). *Monte carlo simulation* (Vol. 116). Sage Publications.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123–137.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Novak, E. (2014). Effects of simulation-based learning on students' statistical factual, conceptual and application knowledge. *Journal of Computer Assisted Learning*, *30*(2), 148–158.

Ören, T. (2009). Uses of simulation. In J. A. Sokolowski & C. M. Banks (Eds.), *Principles of modeling and simulation: A multidisciplinary approach* (pp. 153–179). John Wiley & Sons.

Ören, T. (2011a). A critical review of definitions and about 400 types of modeling and simulation. *SCS M&S Magazine*, *2*(3), 142–151.

Ören, T. (2011b). The many facets of simulation through a collection of about 100 definitions. *SCS M&S Magazine*, *2*(2), 82–92.

Pablo, N., & Chance, B. (2018). Can a simulation-based inference course be flipped? In M.A. Sorto (Ed.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS-10)*, Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM*, 1–11.

Pfannkuch, M., & Budgett, S. (2014). Constructing inferential concepts through bootstrap and randomization-test simulations: A case study. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute. http://icots.info/icots/9/proceedings/pdfs/ICOTS9_8J1_PFANNKUCH.pdf

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, *4*(1), 119–130.

Pruim, R., Kaplan, D. T., & Horton, N. J. (2017). The mosaic Package: Helping Students to "Think with Data" Using R. *The R Journal*, *9*(1), 77–102.

R Core Team. (2019). *R: A language and environment for statistical computing* (3.6.2) [Computer software]. R Foundation for Statistical Computing.

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0.2) [Computer software]. R Foundation for Statistical Computing.

Ramsey, F., & Schafer, D. (2012). *The Statistical Sleuth: A Course in Methods of Data Analysis* (3rd ed.). Cengage Learning.

Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of p-values. *Statistics Education Research Journal*, *13*(1). http://iase-web.org/documents/SERJ/SERJ13(1)_Reaburn.pdf

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research* (1.9.12) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych

Revelle, W., & Condon, D. M. (2019). Reliability from [alpha] to [omega]: A tutorial. *Psychological Assessment*, *31*(12), 1395.

Ricketts, C., & Berry, J. (1994). Teaching statistics through resampling. *Teaching Statistics*, *16*(2), 41–44.

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

Roy, S., & Mcdonnel, T. (2018). Assessing simulation-based inference in secondary schools. In M.A. Sorto (Ed.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS-10)*, Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257–270.

Santrock, J. W. (2011). *Educational Psychology* (5th ed.). McGraw-Hill.

Saputra, K. V., & Couch, O. (2018). Implementing a simulation-based inference curriculum in Indonesia: A preliminary report. In M.A. Sorto (Ed.), *Proceedings*

of the Tenth International Conference on Teaching Statistics (ICOTS-10), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475–5223.

Simon, J. L., Atkinson, D. T., & Shevokas, C. (1976). Probability and statistics: Experimental results of a radically different teaching method. *The American Mathematical Monthly*, *83*(9), 733–739.

Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, *3*(1), 3–9.

Sokolowski, J. A., & Banks, C. M. (Eds.). (2009). *Principles of modeling and simulation: A multidisciplinary approach*. John Wiley & Sons.

Stephens, M., Carver, R., & McCormack, D. (2014). From Data to Decision-Making: Using Simulation and Resampling Methods to Teach Inferential Concepts. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute. http://icots.info/9/proceedings/pdfs/ICOTS9_8B3_STEPHENS.pdf

Swaak, J., de Jong, T., & van Joolingen, W. R. (2004). The effects of discovery learning and expository instruction on the acquisition of definitional and intuitive knowledge. *Journal of Computer Assisted Learning*, *20*(4), 225–234.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296.

Tabor, J., Starnes, D. S., Yates, D. S., & Moore, D. S. (2012). *The practice of statistics: Annotated teacher's edition.* W.H. Freeman.

Taylor, L., & Doehler, K. (2015). Reinforcing Sampling Distributions through a Randomization-Based Activity for Introducing ANOVA. *Journal of Statistics Education*, *23*(3).

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2016). *Introduction to Statistical Investigations* (1st ed.). Wiley.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2018). Assessing the Association Between Precourse Metrics of Student Preparation and Student Performance in Introductory Statistics: Results from Early Data on Simulation-Based Inference vs. Nonsimulation-Based Inference. *Journal of Statistics Education*, *26*(2), 103–109.

Tintle, N. L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute. http://digitalcollections.dordt.edu/faculty_work/64/?utm_source=digitalcollections.dordt.edu%2Ffaculty_work%2F64&utm_medium=PDF&utm_campaign=PDFCoverPages

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.

Tintle, N., & VanderStoep, J. (2018). Development of a tool to assess students' conceptual understanding in introductory statistics. In M.A. Sorto (Ed.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS-10)*, Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

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), n1.

van Dijke-Droogers, M., Drijvers, P., & Bakker, A. (2021). Statistical modeling processes through the lens of instrumental genesis. *Educational Studies in Mathematics*, 1–26.

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 (Ed.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS-10)*, Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.

Weir, C. G., McManus, I. C., & Kiely, B. (1991). Evaluation of the teaching of statistical concepts by interactive experience with Monte Carlo simulations. *British Journal of Educational Psychology*, *61*(2), 240–247.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wood, M. (2005). The role of simulation approaches in statistics. *Journal of Statistics Education*, *13*(3), 1–11.

Ziegler, L. A. (2014). *Reconceptualizing statistical literacy: Developing an assessment for the modern introductory statistics course* [Doctoral dissertation]. https://conservancy.umn.edu/handle/11299/165153

Appendix A: Characteristics of Quantitative Empirical Studies

Appendix A1: Attributes of studies comparing multiple curricula

Table 25 *Attributes of multi-class comparison studies*

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
Tintle, VanderStoep, Holmes, Quisenberry, & Swanson (2011)	- prelim-ISI[a] - CAOS national sample[b] - non-SBI[c]	- 202 - 763 - 195	- CAOS[b]	
Garfield, delMas, & Zieffler (2012)	- CATALST - GOALS/CAOS national sample[d]	- 102 - 5362	- GOALS/CAOS[d]	
Tintle, Topliff, VanderStoep, Holmes, & Swanson (2012)	- prelim-ISI[e] - non-SBI[c]	- 76 - 79	- CAOS[b]	- Focus on 4-month retention - Subsample from Tintle et al. (2011)
Tintle et al., (2014)*	- prelim-ISI[f] - non-SBI[g]	- 155 - 94	- CAOS[b]	
Chance, Wong, & Tintle (2016)	- ISI[h] Instructor experience groups: - Experienced - Middle - New - Non-user - CAOS national sample[b]	- 1116 - 763	- Modified-CAOS[s]	- Comparison results were not the primary focus of analysis - Two “non-user” instructors did not implement ISI throughout the entire time period of study, and thus represent inconsistent instructors regarding group classification

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
Beckman, delMas, & Garfield (2017)	- CATALST[i], University of Minnesota - CATALST[i], other universities - non-SBI[j]	- 138 - 151 - 440	- CATALST custom assessment[t] - Non-CATALST custom assessment[t]	- Focus on comparing cognitive near transfer and far transfer; see paper for descriptions of transfer types
Chance, Mendoza, & Tintle (2018)	- ISI[k], instructors new to textbook - ISI[k], instructors experienced with textbook - Lock 5[l] - CATALST[i] - non-SBI[j], GAISE-compliant - non-SBI[j], non-GAISE-compliant - Other[m]	Approximate total sample size: 15000 (sample sizes not given for all groups)	- Modified-CAOS[s]	
Hildreth, Robison-Cox, & Schmidt (2018)	- CATALST[i] - Lock 5 - Modified Lock 5[o] - non-SBI[p]	- $n_1 = 770$; $n_2 = 699$ - $n_1 = 758$, $n_2 = 768$ - $n_1 = 1500$; $n_2 = 1471$ - $n_1 = 84$; $n_2 = 553$	- Modified-CAOS[u] - Course grade	- Sample sizes vary by analysis: n_1 corresponds to "student understanding" analysis, n_2 to "student success" analysis
Mendoza & Roy (2018)	- ISI[q] - Other SBI[r] - non-SBI[j]	- $n_1 = 197$; $n_2 = 61$ - $n_1 = 127$; $n_2 = 26$ - $n_1 = 284$; $n_2 = 61$	- Modified-CAOS[s]	- Focus on retention - Sample sizes vary by analysis: n_1 corresponds to 4-month retention analysis, n_2 to 16-month retention analysis
Roy & Mcdonnel (2018)	- ISI[k] - non-SBI[j]	- 196 - 435	- Modified-CAOS[s]	- Focus on secondary schools
Saputra & Couch (2018)	- SBI[f] - non-SBI[j]	- 88 - 42	- Modified-CAOS[s]	- Focus on an Indonesian university
Tintle et al., (2018)*	- ISI[a,e,f] - non-SBI[j]	- $n_1 = 366$; $n_2 = 152$; $n_3 = 141$ - $n_1 = 289$; $n_2 = 91$; $n_3 = 80$	- CAOS[b]	- Data originally from Tintle et al. (2011), Tintle et al. (2012), and Tintle et al. (2014) - Focus on sample stratification by student performance variables

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
				- Sample sizes vary by stratification or analysis approach: n_1 corresponds to analysis stratified by pre-course performance, n_2 to ACT score, n_3 to analysis of only low performing students stratified by assessment topic
VanderStoep, Couch, & Lenderink (2018)	- SBI[r] - non-SBI[j]	- $n_1 = 3340$; $n_2 = 2636$; $n_3 = 2766$; $n_4 = 886$ - $n_1 = 1926$; $n_2 = 1619$; $n_3 = 1670$; $n_4 = 601$	- Modified-CAOS[v]	- Expansion of Tintle et al. (2018) - Sample sizes vary by stratification or analysis approach: n_1 corresponds to analysis stratified by pre-course performance, n_2 to SAT/ACT z-score, n_3 to self-reported college GPA, n_4 to analysis of only low performing students stratified by assessment topic

Notes.

*Studies contained additional curricular groups without a comparison group, and thus such groups were not listed.

[a]Preliminary curriculum to ISI (Tintle et al., 2016). [b]delMas, et al. (2007). [c]Agresti and Franklin (2008). [d]Scores from GOALS assessment items comparable to CAOS items, with 2009-2011 CAOS national sample. [e]Developed version of curriculum used in Tintle et al., (2011). [f]2014 version of ISI (Tintle et al., 2016). [g]Unknown non-SBI textbook. [h]2015 version of ISI (Tintle et al., 2016). [i]Garfield, delMas, and Zieffler (2012). [j]Unknown mixture of courses and textbooks. [k]Tintle et al. (2016). [l]2012 version of Lock et al., (2013). [m]Unknown mixture of SBI and non-SBI courses. [n]With adjustments to align topic order with ISI (Tintle et al., 2016). [o]Montana State University version of Lock et al., (2013). [p]de Veaux et al. (2008). [q]Unclear which versions of ISI were used. [r]Unclear which SBI curricula were used. [s]30-item version of CAOS (delMas et al., 2007). [t]16 CAOS items plus five new items. [u]Six items similar to or exactly the same as CAOS items. [v]Exact assessment unclear, as authors point to 24-item version in Tintle and VanderStoep (2018), but study results reference 35 items.

Appendix A2: Attributes of other studies

Table 26 *Attributes of within-class, multi-section, single unit, and isolated intervention studies*

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
Simon, Atkinson, & Shevokas (1976)	Part 1 (single course with no comparison) - Non-simulation statistics course with simulation unit	- 25	Part 1 - 10-item course exam	- Single class non-comparison and multi-class comparison of for an apparent single class unit on statistics and probability - Part 1 focus on whether students chose Monte Carlo methods or analytic methods on an exam when given the option - Parts 2 and 3 focus on attitudes and learning
	Part 2 (three concurrent classes with a statistics and probability unit) - Non-simulation - Simulation with computer - Simulation without a computer	 - 19 - 39 - 13	Part 2 - 7-item course exam	
	Part 3 (two semesters of two concurrent classes with a varying second-half of each course) - Two simulation groups - Two non-simulation groups	 - 58 - 55	Part 3 - 3 or 4 course exam items	
Weir, McManus, & Kiely (1991)	Part 1 - Simulation for standard error of the mean - Simulation for F-statistic in ANOVA	 - 20 - 19	Both parts - Course exams (unclear number of items and item types, though some open-ended)	- Unclear study design - Apparent within-class comparison of different learning topic groups (standard error of a mean or F-statistic), for part of a course - In both parts, students assigned to matched groups based on course achievement from a prior course (unclear of the extent to which random assignment involved) - Focus on understanding sampling distributions, the F-statistic in ANOVA, and attitudes - Outcomes focus on extent to which students perform well on problems
	Part 2 - Simulation for standard error of the mean - Simulation for F-statistic in ANOVA - Control who watched simulation demonstration	 - 23 - 21 - 21		

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
				similar to their learning condition and different form their learning condition - Part 2 used adjusted materials compared to Part 1
delMas, Garfield, & Chance (1999)	Pilot: - Two non-simulation courses with a simulation component Part 2 (Initial Activity) - Two non-simulation courses with a simulation component (total completing pre- and post-tests, n = 89) Part 3 (New Activity) - Three non-simulation courses with a simulation component (total completing pre- and post-tests, n = 141)	- 51 (total) - 89 (total completing pre- and post-tests) - 141 (total completing pre- and post-tests)	Pilot: - 5 multi-part open-ended items based on visual assessment of distributions Parts 2 and 3 - Updated version of pilot assessment	- Within-class study; no comparison groups - Focus on a model for classroom research - Focus on using a simulation activity for sampling distribution understanding, with assessment of visuals - Adjusted simulation, activity, and assessment for each of three parts - New activity focused on exposing students to having a contradictory experience
Lane & Tang (2000)	- Computer simulation and specific training questions (Sim / Specific) - Computer simulation and non-specific training questions (Sim / Non-specific) - Textbook and specific training questions (Text / Specific) - Textbook and non-specific training questions (Text / Non-specific) - No training (Control)	Total sample size: 115 (group sizes unclear)	- 12 open-ended items adapted from Chen (1999)	- Randomized experiment - One-time isolated intervention - Simulation group watched a video with an accompanying audio script playing in the background - Textbook group read about concepts in a textbook - Conceptual purpose to understand the effect of sample size and on a sampling distribution - Focus on transfer to real-world situations
Mills (2004)	- Simulation with Microsoft Excel - Non-simulation	- 14 - 17	- 7 open-ended items (pre- and post-tests)	- Somewhat unclear study design; comparison of volunteer students from

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
			- 5 multiple choice items as part of course exam	the same course randomly assigned to two groups for a learning and testing intervention outside of class - Focus on conceptual development of Central Limit Theorem, attitudes, and performance on subsequent course exam - Groups only differed by how concepts were practiced, while content was the same
Francis, Kokonis, & Lipson, (2007)	- First simulation exposure before content lecture - First simulation exposure after content lecture	- 43 - 33	- 4 items (2 items similar to each other on two different course exams)	- Unclear study design; within-class comparison of two groups of students with unclear assignment and unclear delivery of an intervention (either separate from or as part of the course for all students) - Focus on sequence of exposure to simulation activities, relative to lecture - Focus on ordered stages for building a hypothesis testing schema - Multiple concepts addressed
Holcomb, Chance, Rossman, Tietjen, et al. (2010)	Part 1 - Given observed data leading to significant outcome (Significant group) - Given observed data leading to non-significant outcome (Non-significant group)	Unclear sample sizes	Part 1 - 3 items totaling 6 prompts (mixture of multiple choice and short answer)	- Multi-section (one class) classroom experiments - Randomized experiment in both parts - Focus on answering inferential questions from real data in both parts
	Part 2 - Manual simulation and computer simulation - Computer simulation	- 20 - 23	Part 2 - 5 multiple choice items focused on one situation	
Novak (2014)	- Storyline integrated with simulation - No storyline with simulation	- 32 - 32	- 18-item parallel pre- and post-tests	- Randomized experiment - One-time isolated intervention

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
				- Focus on learning enhanced with storylines paired with simulations - Emphasis on descriptive statistics topics
Reaburn (2014)	- Pre-intervention (first semester) - First intervention (second semester) - Second intervention (third semester) - Third intervention (fourth semester)	- 12 - 23 - 6 - 12	- 2 open-ended items from an end of semester test	- Multi-year (one class) comparison - Course-long intervention, with subjects volunteering their data - Focus on p-value understanding - One introductory course taught four times in succession, with the course being updated based on experience from the previous iteration - Pre-intervention: no simulation - First intervention (second semester): addition of informal inferential reasoning and some interactive computer simulations - Second intervention (third semester): contextual p-value logic example added - Third intervention (fourth semester) introduced additional p-value logic via diagrams and philosophy of science
Taylor & Doehler (2015)	- Simulation introduction to ANOVA - Non-simulation introduction to ANOVA	- 38 (two-semester total) - 40 (two-semester total)	- Open-ended parallel pre- and post-tests (5 and 6 items, respectively) developed for this study	- Multi-section comparison of a topic unit in the same course (two instructors taught the same section twice) - One section for each instructor randomly assigned to the simulation approach to given topic unit - Focus on understanding ANOVA F-test and sampling distributions with randomization test - Groups only differ by activity and introduction to ANOVA topic, while subsequent content was the same

Study	Comparison Groups	Sample Sizes	Learning Assessment(s)	Design Comments
Maurer & Lock (2016)	- Simulation curriculum past week 9 using StatKey and similar to Lock et al. (2013) - Non-simulation curriculum past week 9 similar to Agresti and Franklin (2012)	- 51 (two-cohort total) - 50 (two-cohort total)	- 20 multiple-choice items from ARTIST[c] used on the final exam - 2 open-ended items from the final exam	- Multi-section comparison of two different curricula for the second-half of each section - Random assignment of students from two sections to each curriculum (including room assignment, schedule, etc.) within each section - At week 9 of course the two curricula diverged in how similar concepts were taught - Focus on confidence intervals and hypothesis testing
Pablo & Chance (2018)	- ISI[a] flipped classroom - ISI[a] non-flipped classroom	- 22 - 19	- Modified-CAOS[b]	- Full course intervention comparison; same instructor for both courses - Focus on comparing effects of a flipped classroom between two simulation sections of the same course

Notes.
[a]Tintle et al. (2016). [b]Modified version of CAOS (delMas et al., 2007) for ISI. [c]Assessment Resource Tools for Improving Statistical Thinking (ARTIST) (Garfield et al., 2006).

Appendix B: Versions of SUSIE

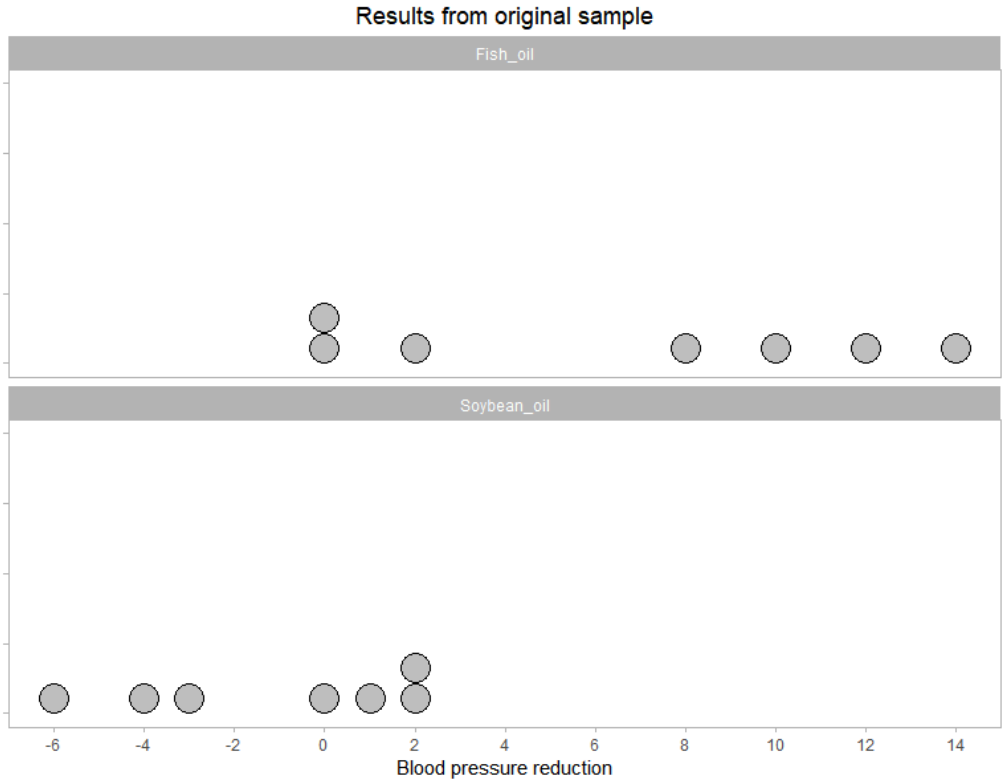
Appendix B1: Initial SUSIE versions for Think-aloud interviews

Initially, four versions of the instrument were created for interviews. Following iterative updates to the instrument throughout the interview process, a total of seven versions were tested. A complete instrument from the initial instrument creation is provided below. This is a combination of Version A and Version B. Version A used "Wording B" (indicated below) for the two items about running more trials (labeled as iN6 and iE6 throughout) and Version B used "Wording A" (indicated below). To show the initial wording variations that were tested, both initial variations on those two items are included. Thus, both iN6 and iE6 appear twice in each instrument section.

Situation 1
Two researchers (called "Researcher 1" and "Researcher 2") investigated whether fish oil can reduce blood pressure more than other types of oils. They recruited fourteen male participants with high blood pressure and randomly assigned each person to one of two treatment groups. The first treatment was a four-week diet that included fish oil. The second was a four-week diet that included soybean oil.

The diastolic blood pressure of each participant was measured twice: once at the beginning and again at the end of the four weeks. The reduction in blood pressure was recorded. Positive values indicate blood pressure went down, and negative values indicate blood pressure went up.

The following research question was proposed: Is there a difference in the average blood pressure reduction between the fish oil group and the soybean oil group?
The study results were plotted, and the sample average difference was calculated:



Sample average difference = Average of fish oil group – average of soybean oil group = 6.6 – (-1.1) = 7.7

To evaluate whether the difference of 7.7 was simply due to random chance, the researchers created a computer simulation that used the following algorithm:

1. Input the original 14 values from the observed data.
2. Randomly assign 7 of those values into one group and the other 7 values into a second group.
3. Label the first group as "Fish oil" and the second group as "Soybean oil".
4. Calculate the average of the 7 values in the Fish oil group and the average of the 7 values in the Soybean oil group.
5. Calculate and record the difference in averages between the two groups.

The computer repeated the process above for a large number of trials.

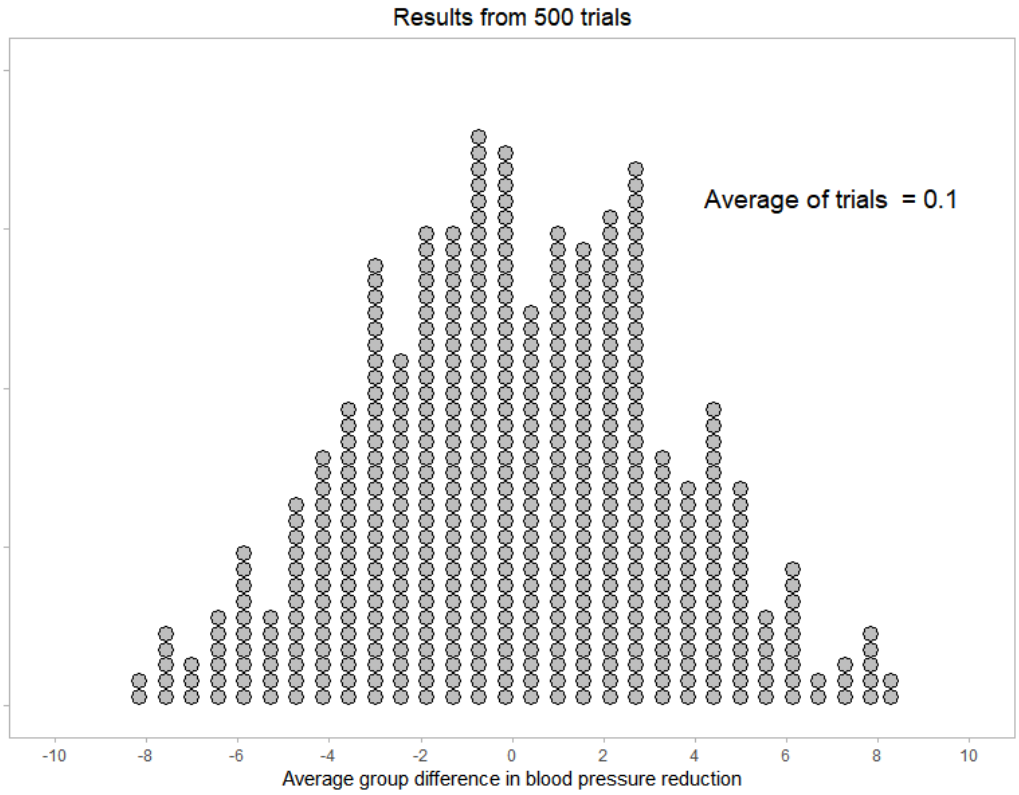
iN1

What assumption about the effect of fish oil relative to soybean oil is the computer simulation designed to model?

iN2

How does the assumption you just discussed help answer the research question of whether there is a difference in the average blood pressure reduction between the two groups?

Researcher 1 plotted the results and calculated the average from the 500 trials, as shown:



iN3

After seeing the distribution of 500 trials, Researcher 1 offered this observation: "Since the results center around 0, that suggests there is no difference in effect between fish oil and soybean oil in reality."
Do you agree with Researcher 1? Why or why not?

iN4

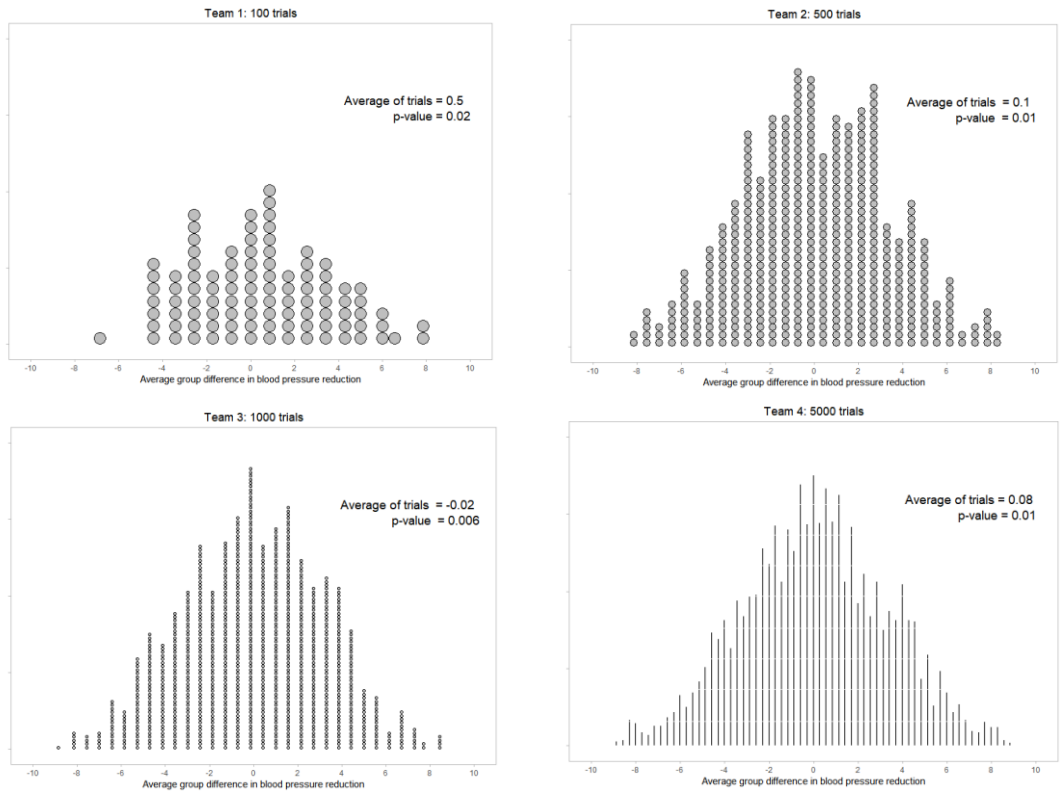
The researchers agreed to calculate a p-value but disagreed on the procedure.
Researcher 1: "Start at the average of the 500 trials, and then calculate the proportion of all trials larger than that."
Researcher 2: "Start at the sample average difference of 7.7, and then calculate the proportion of all trials larger than that."
Which researcher do you agree with? Why?

iN5

Researcher 2 calculated a p-value of 0.014. This indicates the difference of 7.7 or larger was unlikely to occur by chance, assuming no difference between the two groups.
However, Researcher 2 was unsure of how to answer the research question.
Possible answer 1: "The low p-value indicates that an average difference of 0 is not supported. Therefore, there is evidence for a difference between the groups."
Possible answer 2: "The low p-value indicates that the sample average difference of 7.7 from the data is not supported. As a result, there is no evidence for a difference between the groups."
Which answer do you agree with? Why?

iN6 WORDING A and WORDING B

Now consider four different teams of researchers, where each team runs the simulation a different number of times. Team 1 uses 100 trials, Team 2 uses 500 trials, Team 3 uses 1000 trials, and Team 4 uses 5000 trials. Each set of results are shown as follows:



WORDING A: Consider the differences in the four sets of results above. Why might changing the number of trials be related to any of these differences?

WORDING B: If you had to choose the results from one team to use, which team would you choose? For the team you chose, why is their number of trials appropriate for answering the research question?

After the analysis was completed, suppose a supervisor discovered three different concerns about the original data.

iN7

Supervisor: "All of those in the fish oil group were taking a blood pressure medication, while none of those in the soybean oil group were taking any blood pressure medication."
Researcher 1: "Since each trial of the simulation randomly assigned the blood pressure values to the two groups, the presence of the medication is not a problem."

Do you agree with the argument from Researcher 1? Why or why not?

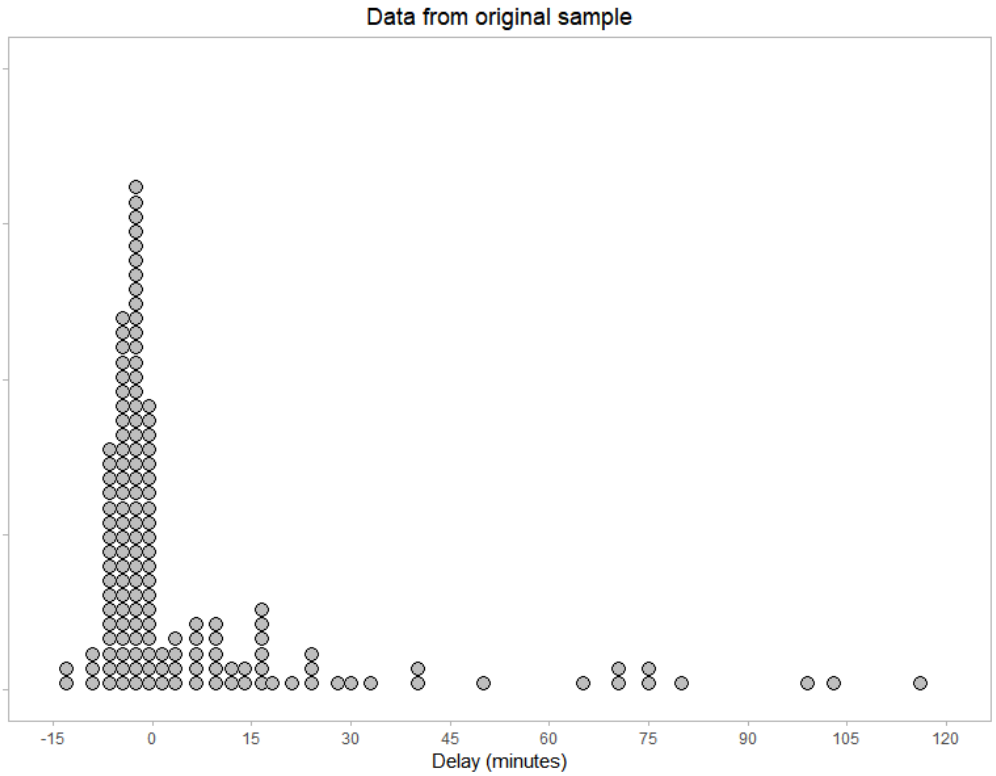
iN8
Supervisor: “The sample size is only 14. This seems too small.”
Researcher 2: “The small sample size is OK, because we ran many trials of the simulation.”
Do you agree with the argument from Researcher 2? Why or why not?

iN9
Supervisor: “The medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have.”
Given the simulation that was run, can the researchers still provide a valid answer to the research question? Why or why not?

iN10
Since this was only one study the researchers wanted to do a replication study, to verify the results they just found. However, they disagreed on what to do.
Researcher 1: “We should recruit 14 new participants and put them through the same study that we just did.”
Researcher 2: “Collecting new participants is unnecessary. We can just run our simulation again, as that is as good as a replication study.”
Which researcher do you agree with? Why?

Situation 2
Airplane delays are a common complaint from travelers. Suppose you have two friends (“Friend 1” and “Friend 2”) who want to know the average delay for recent departures from Minneapolis-St. Paul airport (MSP).
Using a US Department of Transportation database, your friends collected a random sample of 150 departure delay times from MSP airport for Delta Airlines in 2019. A positive number indicates departing late, and a negative number indicates departing early. The following research question was proposed: What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

The sample of 150 values was plotted, and the sample average was calculated:



Sample average delay = 6.35 minutes

To estimate the uncertainty in the average of 6.35, your friends created a computer bootstrap simulation that used the following algorithm:

1. Input the original 150 values from the observed data.
2. Randomly draw one value, record it, and put it back in the list.
3. Repeat Step 2 until 150 values are recorded.
4. Calculate the average delay time from the 150 recorded values.

The computer repeated the process above for a large number of trials.

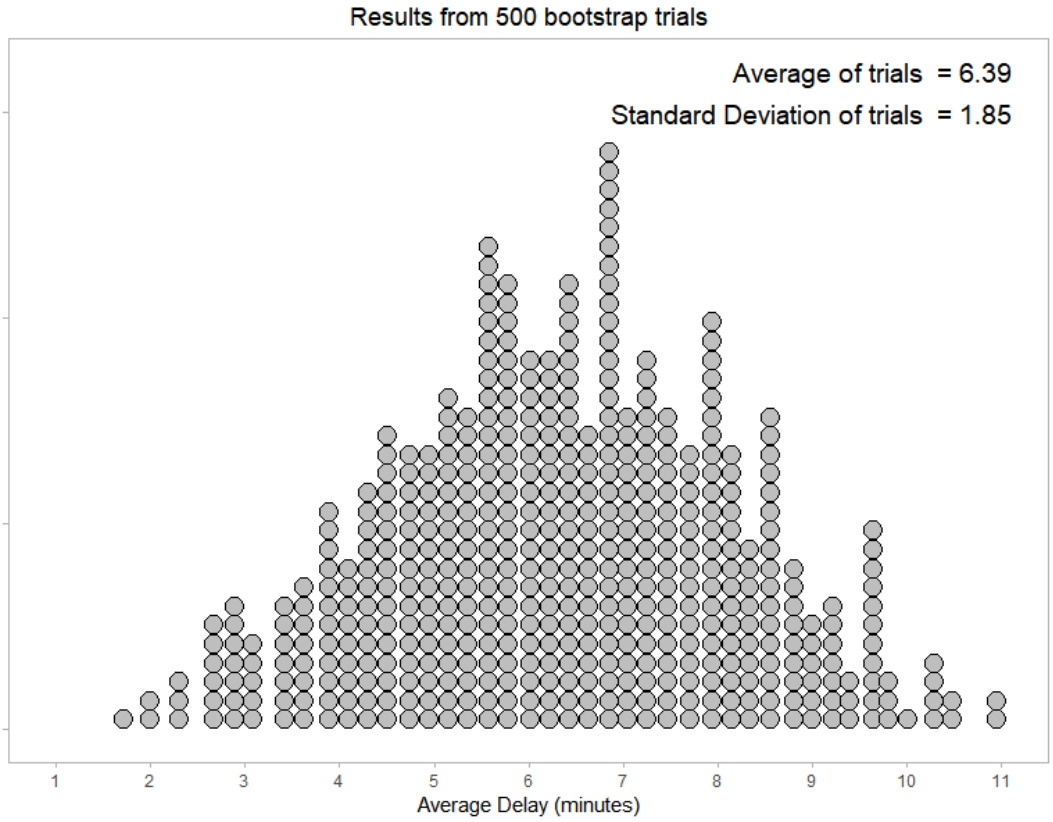
iE1

What assumption about the sample and its variability is the computer simulation designed to model?

iE2

How does the assumption you discussed in Question # help answer the research question about the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

Friend 2 plotted the results and calculated the average and standard deviation from the 500 trials, as shown:



Next, your friends calculated an interval for the average delay time using the following formula:

Interval = *Sample Estimate* ± 2 * *Standard Deviation of trials.*

iE3

Your friends agreed to use 1.85 for the *Standard Deviation of trials* but disagreed on the *Sample Estimate*.

Friend 1: "The sample average of 6.35 minutes should be used."

Friend 2: "The average of the 500 bootstrap trials, 6.39 minutes, should be used."

Which friend do you agree with? Why?

iE4

Friend 1 calculated an interval of [2.65, 10.05]. However, Friend 2 argued that the interval was not needed to answer the research question.

Friend 2: "We can just use the average of the 500 bootstrap trials. We can be certain that the average flight delay for all departures from MSP with Delta Airlines in 2019 was 6.39 minutes."

Do you agree with this statement? Why or why not?

iE5

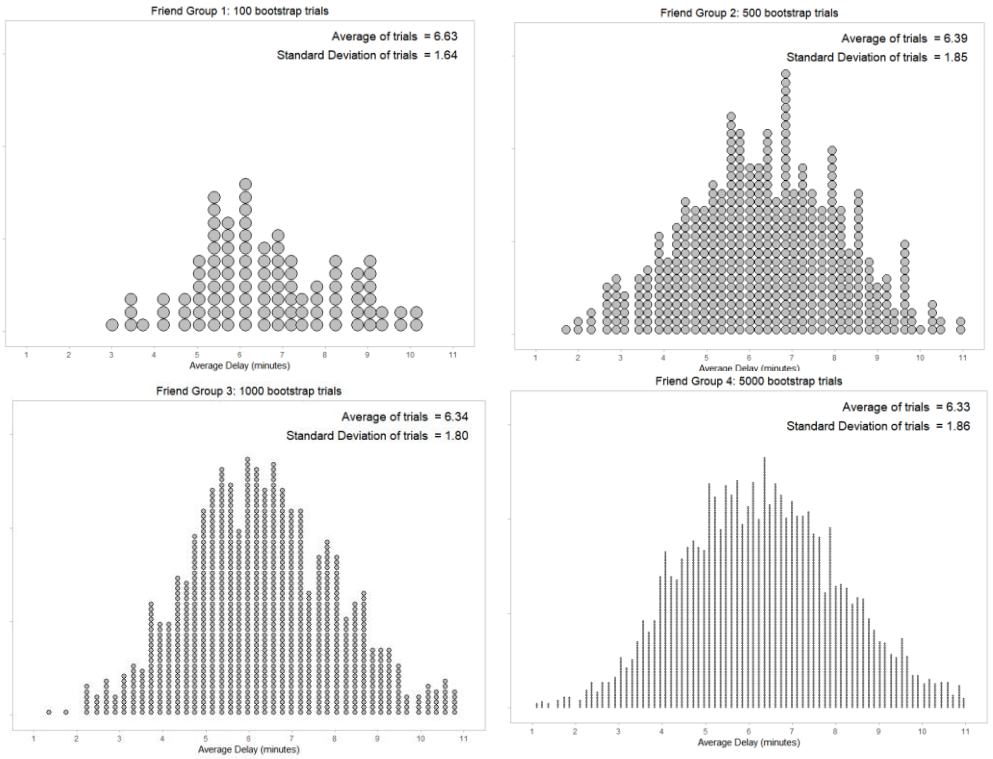
After the analysis was completed, suppose a third friend saw the plot of 500 trials and raised a concern.

Friend 3: "The average of 6.39 from the 500 bootstrap trials is different from the sample average of 6.35. This indicates we should not trust the sample data nor the sample average of 6.35."

Do you agree with this statement? Why or why not?

iE6 WORDING A and WORDING B

Now consider four different groups of friends, where each group runs the bootstrap simulation a different number of times. Group 1 uses 100 trials, Group 2 uses 500 trials, Group 3 uses 1000 trials, and Group 4 uses 5000 trials. Each set of results is shown as follows:



WORDING A: Consider the differences in the fours sets of results above. Why might changing the number of trials be related to any of these differences?

WORDING B: If you had to choose the results from one friend group to use, which friend group would you choose? For the friend group you chose, why is their number of trials appropriate for answering the research question?

After the analysis was completed, suppose your friend's statistics teacher discovered three different concerns about the original data.

iE7

Teacher: "Only morning departure times (in the range of 5am – 12pm) were included in the sample."

Given the bootstrap simulation that was run, can your friends still provide a valid answer to the research question? Why or why not?

iE8
Teacher: “The sample size is only 150. This seems too small.”
Friend 1: “The small sample size is OK, because we ran many trials of the bootstrap simulation.”
Do you agree with the argument from Friend 1? Why or why not?

iE9
Teacher: The flight tracker at MSP airport was found to be faulty for Delta airlines. It consistently recorded departure times as being much later than it should have.
Given the bootstrap simulation that was run, can your friends still provide a valid answer to the research question? Why or why not?

iE10
Since this was only one sample of data, your friends wanted to do a replication study to verify the results they just found. However, they disagreed on what to do.
Friend 1: “Collecting more data is unnecessary. We can just run our bootstrap simulation again, as that is as good as a replication study.”
Friend 2: “We should collect another sample of 150 departures and analyze it the same way that we analyzed the original sample.”
Which friend do you agree with? Why or why not?

Appendix B2: Final version of SUSIE for field test

The final instrument for the field test followed from the feedback and edits applied throughout the interviews. Field test participants were requested to answer questions for both the fish oil section and the airplane delays section, but each participant was assigned the sections in a random order. Of these two possible versions of the instrument, only the instrument starting with the fish oil section is shown below. The instrument is shown as a series of screenshots as participants saw them in Qualtrics. Each screenshot page is labeled.

Following the last page of this instrument, participants were automatically redirected to a separate Qualtrics survey asking for their email and class section. That information was used to notify instructors which of their students should receive extra credit. That survey is not shown here.

For clarity, the item designations from the instrument blueprint were superimposed on each item below, e.g., [Item n01]. Participants did not see these item designations.

Page 1

Student Understanding of Statistical Simulations

Welcome!

This test consists of two sections. When you are done answering all questions on a given page, click the arrow in the lower-right corner to advance to the next page. Note: You CANNOT go back to earlier pages.

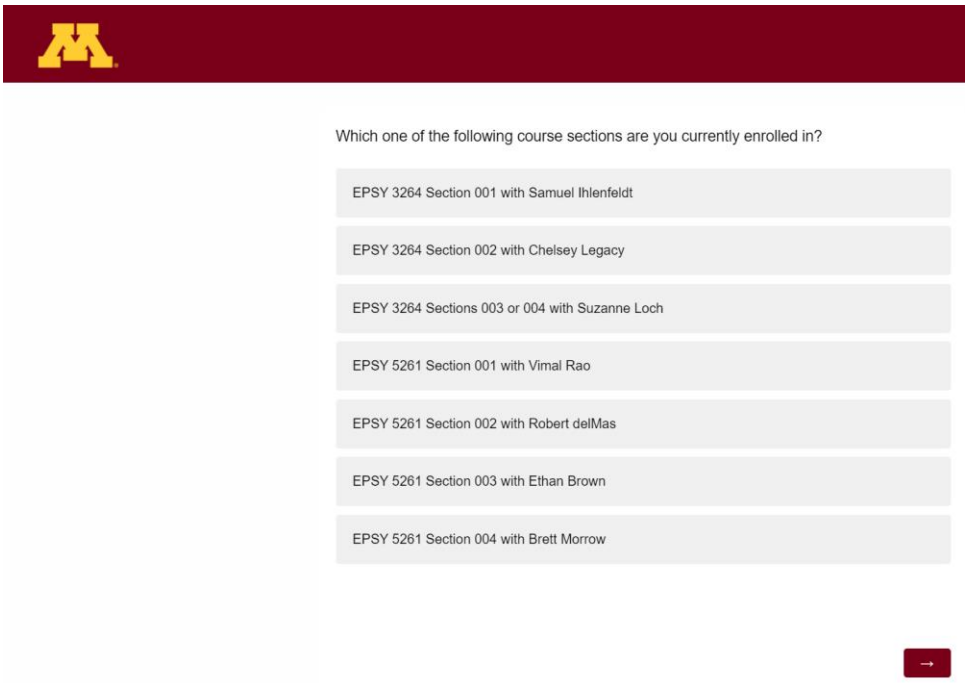
Once you complete the test, you will be taken to a separate Qualtrics survey where you must indicate your class section and email, to receive the extra credit.

The Information Sheet about the study was emailed to you; it is also provided below. If you did not review this sheet yet, please do so before proceeding.

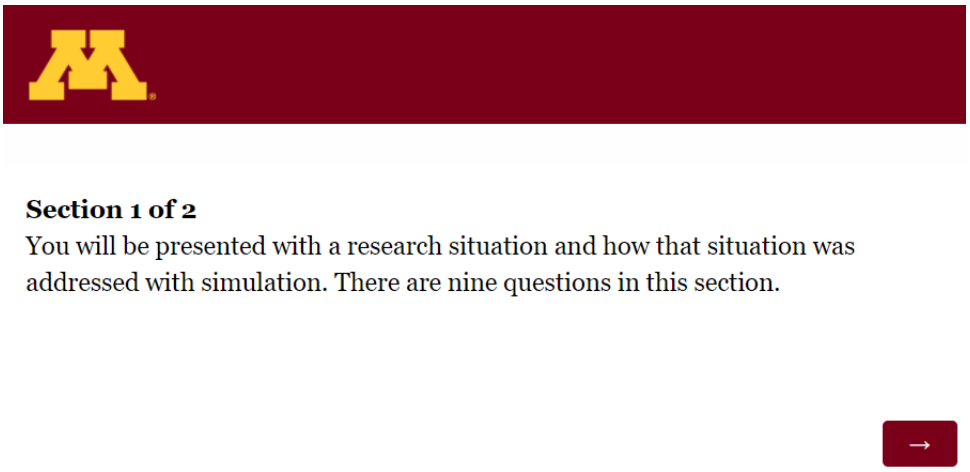
Information Sheet

Please proceed when ready!

Page 2


Which one of the following course sections are you currently enrolled in?
EPSY 3264 Section 001 with Samuel Ihlenfeldt
EPSY 3264 Section 002 with Chelsey Legacy
EPSY 3264 Sections 003 or 004 with Suzanne Loch
EPSY 5261 Section 001 with Vimal Rao
EPSY 5261 Section 002 with Robert delMas
EPSY 5261 Section 003 with Ethan Brown
EPSY 5261 Section 004 with Brett Morrow
→

Page 3


Section 1 of 2
You will be presented with a research situation and how that situation was addressed with simulation. There are nine questions in this section.
→

Situation: Fish Oil and Blood Pressure

Two researchers (called "Researcher 1" and "Researcher 2") investigated whether fish oil can reduce blood pressure more than other types of oils. They randomly sampled fourteen participants from the population of interest and randomly assigned each person to one of two treatment groups. The first treatment was a four-week diet that included fish oil. The second was a four-week diet that included soybean oil.

The diastolic blood pressure of each participant was measured twice: once at the beginning and again at the end of the four weeks. The reduction in blood pressure was recorded. Positive values indicate blood pressure went down, and negative values indicate blood pressure went up.

Research Question

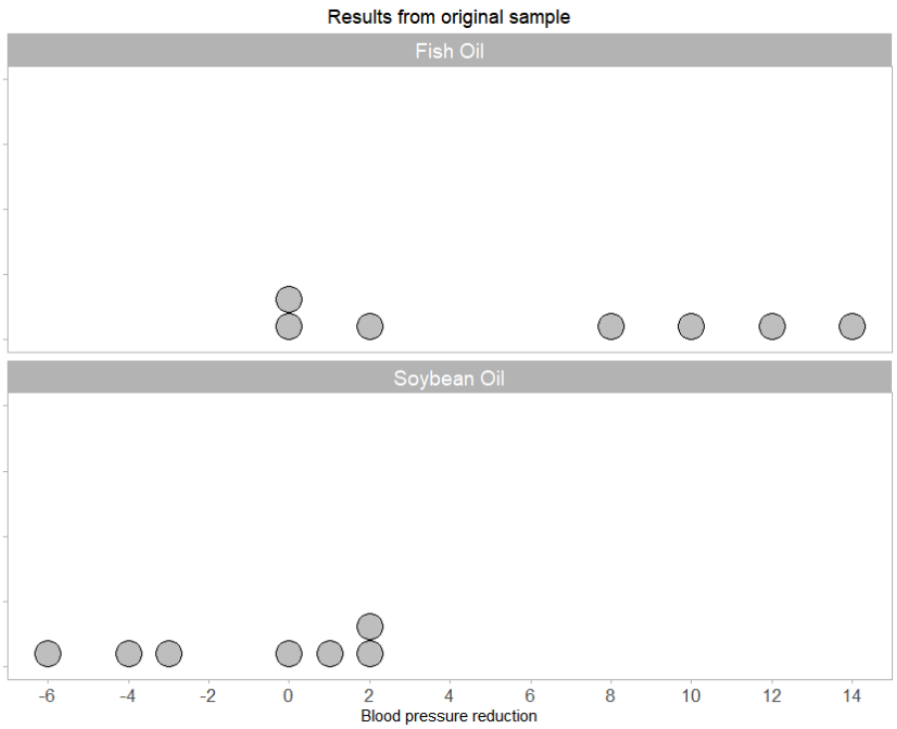
Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?

Data

The sample average difference was calculated as follows:

Average of fish oil group – average of soybean oil group = 6.6 – (-1.1) = **7.7**

The study results were plotted, as shown below. (Note: You do not need to memorize this graph.)



(There are no questions on this page. Please proceed when you are done reading.)

→

(reminder of research question)

Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?

Simulation

To evaluate whether the difference of 7.7 was simply due to random chance, the researchers created a computer simulation that used the following algorithm:

1. Input the original 14 values from the observed data.
2. Randomly assign 7 of those values into one group and the other 7 values into a second group.
3. Label the first group as "Fish oil" and the second group as "Soybean oil".
4. Calculate the average of the 7 values in the Fish oil group and the average of the 7 values in the Soybean oil group.
5. Calculate and record the difference in averages between the two groups.

The computer repeated the process above for a large number of trials.

[Item n01]

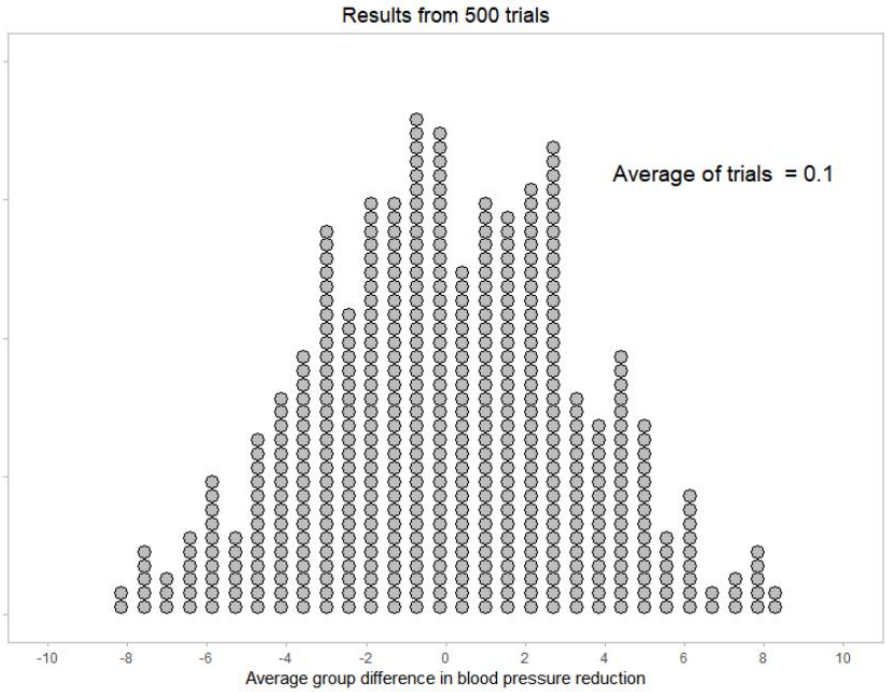
QUESTION

What is the purpose of randomization in the simulation (in terms of how it helps you answer the research question)?

→

Results

Researcher 1 plotted the results and calculated the average from 500 trials of the simulation, as shown:



[Item n02]

QUESTION

After seeing the distribution of 500 trials, Researcher 1 offered this observation:

"Since the results center around 0, that suggests there is no difference in effect between fish oil and soybean oil in reality."

Do you agree with Researcher 1? Why or why not?

[Item n03]

QUESTION

The researchers agreed to calculate a p-value but disagreed on the procedure.

Researcher 1: *"Start at the average of the 500 trials, and then calculate the proportion of all trials larger than that."*

Researcher 2: *"Start at the sample average difference of 7.7, and then calculate the proportion of all trials larger than that."*

Which researcher do you agree with? Why?

→

Page 7

(reminder of research question)

Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?

[Item n04]

QUESTION

Researcher 2 calculated a p-value of 0.014. However, Researcher 2 was unsure of how to answer the research question.

Possible answer 1: *"The low p-value indicates that an average difference of 0 is not supported. There is evidence for a difference between the groups."*

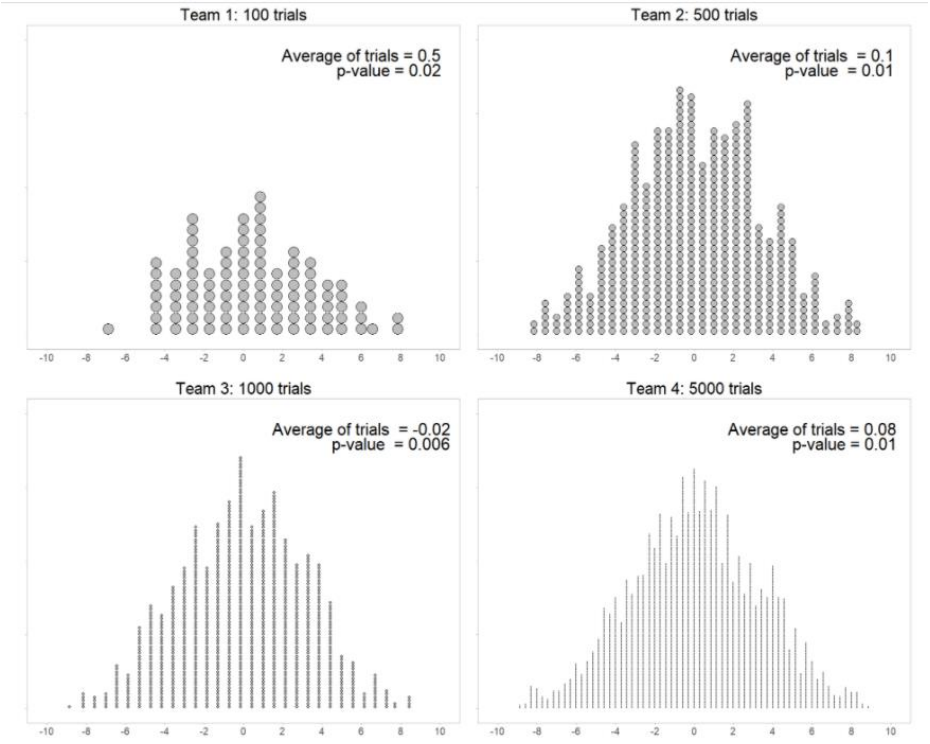
Possible answer 2: *"The low p-value indicates that the sample average difference of 7.7 from the data is not supported. There is no evidence for a difference between the groups."*

Which answer do you agree with? Why?

→

Page 8

Now consider four different teams of researchers, where each team runs the simulation a different number of times. Team 1 uses 100 trials, Team 2 uses 500 trials, Team 3 uses 1000 trials, and Team 4 uses 5000 trials. Each set of results are shown as follows:



[Item n05]

QUESTION

Consider the results from the smallest number of trials (100). Was there something gained by the other teams running more trials? If so, what did they gain and why? If nothing was gained, why not?

Page 9

(reminder of research question)

Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?

[Item n06]

QUESTION

Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results.

Researcher 1: *"We should randomly sample another 14 participants and put them through the same study that we just did."*
Researcher 2: *"We can also run our simulation again with the data we already have, as that is as good as a replication of the study."*

Is the approach by Researcher 2 a valid way to do a replication study? Why or why not?

→

Page 10

Follow-up

The next three questions ask you to consider three hypothetical issues about the data. When answering each question, only consider the ONE specific issue discussed in the given question.

→

(reminder of research question)

Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?

[Item n07]

QUESTION

Suppose that all of those in the fish oil group were taking a blood pressure medication, while all of those in the soybean oil group were not.

Since each trial of the simulation randomly assigned the blood pressure values to the two groups, could the researchers still provide a valid answer to the research question? Why or why not?

→

(reminder of research question)

Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?

[Item n08]

QUESTION

Suppose there was concern that the size of the original sample was too small.

Did the fact that the researchers ran many trials of the simulation create a larger sample size for the study? Why or why not?

→

Page 13

(reminder of research question)

Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?

[Item n09]

QUESTION

Suppose that the medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have.

Given the simulation that was run, could the researchers still provide a valid answer to the research question? Why or why not?

→

Page 14

Section 2 of 2

You will be presented with a different research situation and how that situation was addressed with simulation. There are nine questions in this section.

→

Situation: Airplane Delays

Airplane delays are a common complaint from travelers. Suppose you have two friends ("Friend 1" and "Friend 2") who want to know the average delay for recent departures from Minneapolis-St. Paul airport (MSP).

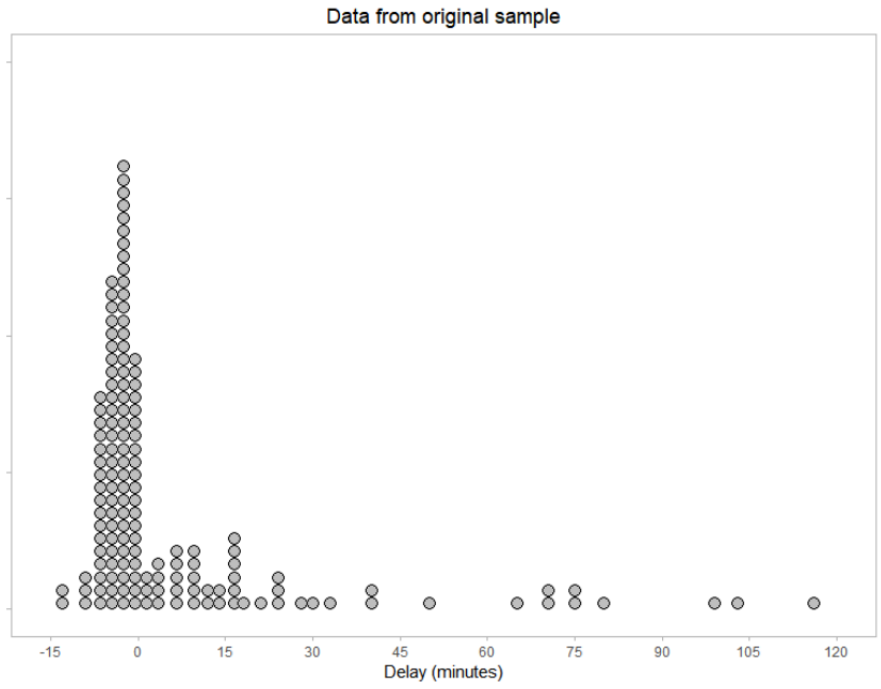
Using a US Department of Transportation database, your friends collected a random sample of 150 departure times from MSP airport for Delta Airlines in 2019. A positive number indicates departing late, and a negative number indicates departing early.

Research Question

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

Data

The sample average delay was calculated to be 6.35 minutes. Also, the sample of 150 values was plotted, as shown below. (Note: You do not need to memorize this graph.)



(There are no questions on this page. Please proceed when you are done reading.)

→

Page 16

(reminder of research question)

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

Simulation

To estimate the uncertainty in the average of 6.35 minutes, your friends created a computer bootstrap simulation that used the following algorithm:

1. Input the original 150 values from the observed data.
2. Randomly draw one value, record it, and put it back in the list.
3. Repeat Step 2 until 150 values are recorded.
4. Calculate the average delay time from the 150 recorded values.

The computer repeated the process above for a large number of trials.

[Item e01]

QUESTION

What is the purpose of resampling in the bootstrap simulation (in terms of how it helps you answer the research question)?

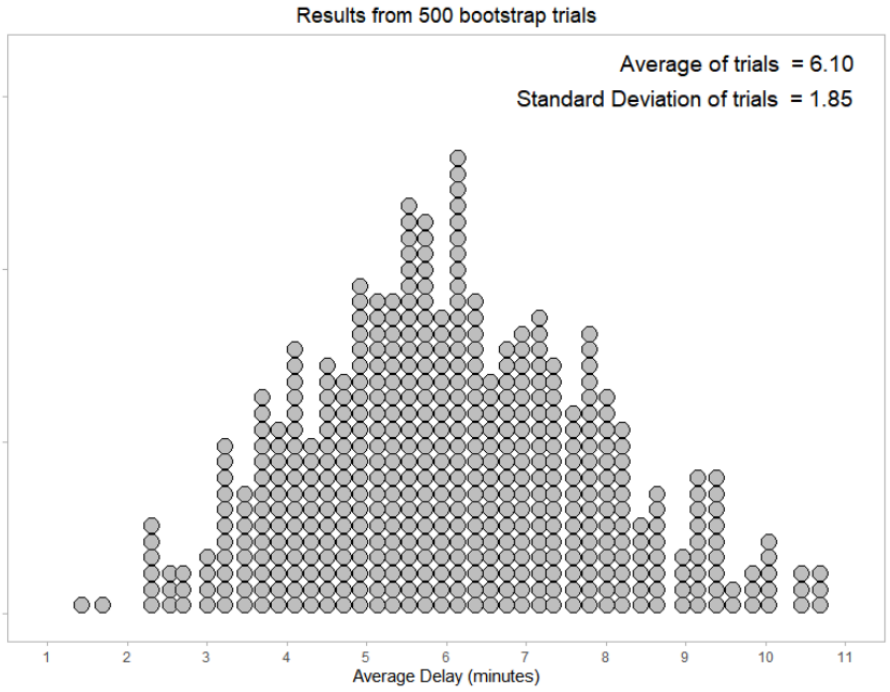
→

(reminder of research question)

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

Results

Friend 2 plotted the results and calculated the average and standard deviation from 500 trials of the simulation, as shown:



Next, your friends calculated a confidence or compatibility interval for the average delay time using the following formula:

Interval = *Sample Estimate* ± 2 * *Standard Deviation of trials*

[Item e02]

QUESTION

Your friends agreed to use 1.85 for the Standard Deviation of trials but disagreed on the Sample Estimate.

Friend 1: *"The sample average of 6.35 minutes should be used."*

Friend 2: *"The average of the 500 bootstrap trials, 6.10 minutes, should be used."*

Which friend do you agree with? Why?

→

(reminder of research question)

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

[Item e03]

QUESTION

To answer the research question, Friend 1 wanted to report the interval they calculated. However, Friend 2 said, *"You don't need an interval to answer the research question. You can report the average of the 500 bootstrap trials."*

Do you agree with Friend 2? Why or why not?

→

Page 19

[Item e04]

QUESTION

Suppose a third friend saw the plot of 500 trials and raised a concern.

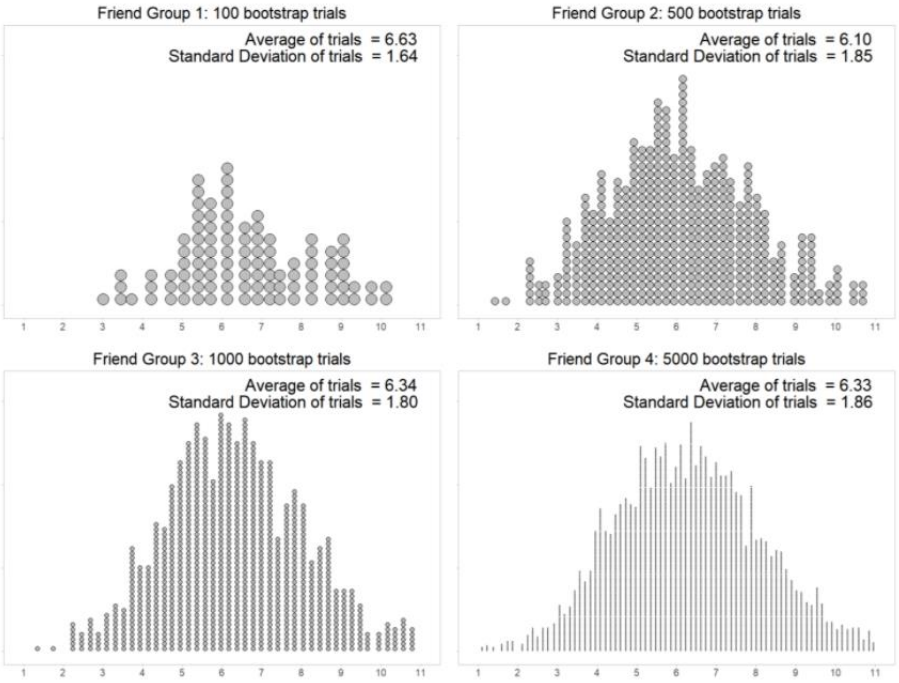
Friend 3: *"The average of 6.10 from the bootstrap trials is different than the sample average of 6.35. When these two averages are different, this suggests the bootstrap average is more trustworthy than the original sample average."*

Do you agree with Friend 3? Why or why not?

→

Page 20

Now consider four different groups of friends, where each group runs the bootstrap simulation a different number of times. Group 1 uses 100 trials, Group 2 uses 500 trials, Group 3 uses 1000 trials, and Group 4 uses 5000 trials. Each set of results is shown as follows:



[Item e05]

QUESTION

Consider the results from the smallest number of trials (100). Was there something gained by the other groups running more trials? If so, what did they gain and why? If nothing was gained, why not?

Page 21

(reminder of research question)

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

[Item e06]

QUESTION

Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results.

Friend 1: *"We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample."*
Friend 2: *"We can also run our bootstrap simulation again with the data we already have, as that is as good as a replication of the study."*

Is the approach by Friend 2 a valid way to do a replication study? Why or why not?

→

Page 22

Follow-up

The next three questions ask you to consider three hypothetical issues about the data. When answering each question, only consider the ONE specific issue discussed in the given question.

→

(reminder of research question)

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

[Item e07]

QUESTION

Suppose that only morning departure times (in the range of 5am – 12pm) were included in the sample.

Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

→

(reminder of research question)

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

[Item e08]

QUESTION

Suppose there was concern that the size of the original sample was too small.

Did the fact that your friends ran many trials of the bootstrap simulation create a larger sample size for the study? Why or why not?

→

Page 25

(reminder of research question)

What is the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

[Item e09]

QUESTION

Suppose that the flight tracker at MSP airport was found to be faulty. It consistently recorded departure times as being later than it should have.

Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

→

Page 26

To submit your responses and to be taken to the extra credit page, please click the arrow in the lower-right corner.

→

Appendix C: Interview Notes and Instrument Changes

Appendix C1: Complete interviewer notes

Unedited interviewer notes are shown for the twelve interviews below. Notes were typed by the interviewer during and after each interview. Comments in single quotes are additional interviewer thoughts regarding instrument changes or side comments about the interviewee.

Interview 01
Overall: ~5 minutes per question; Only made it through end of NHST; all items seem to get at intended idea, though sample size item seemed to bring in distracting issues;

- Word "diastolic"
- Sorting between negative and positive sample values in fish oil situation
- Recognition of answer on NHST Q1: ~5 minutes; INTENDED idea
- LONG reasoning process on NHST Q2: ~ 5 minutes; INTENDED idea, just a long response process
- NHST Q3: Trying to bring in the observed result of 7.7 already, even though that hasn't been incorporated yet in this question; INTENDED idea, with a long answer that technically answered the next questions; ~ 5 minutes
- NHST Q4: relied on "definition of p-value"; tried to explain Research 1's perspective, but couldn't; CORRECT answer, not necessarily intended idea; ~ 5 minutes
- NHST Q5: INTENDED idea? ('this question may be redundant with Q4 for the purposes of this instrument') had the correct idea with Possible answer 1, but was trying to understand the logic of Possible answer 2; tried to reconcile why the wrong answer was proposed; ~ 5 minutes
- NHST Q6: sample size vs. trials; spent long time deliberating between these pictures; but in the end, seemed to get INTENDED idea!
- NHST Q7: was able to separate simulation from real experiment; BUT, had to ask to verify that the random assignment was used in the original experiment, which is not stated in the question; INTENDED idea; drawn-out reasoning process, as intended, but time-consuming
'Change to consider: change Researcher 1's assertion to "…it should not affect the conclusion."'
- NHST Q8: INTENDED idea at first; brought in other issues about sampling and where the sample came from; external validity; focused on the PROBLEM of a small sample size, and less on the trials vs. sample size, though they did recognize it
'Consider change: "Do you agree that running more trials of the simulation can overcome a small sample size?"'
- NHST Q9: INTENDED idea; did focus on how the simulation was one thing and the original data were another;
- NHST Q10: INTENDED idea;

- Overall: hadn't thought about some researchers reactions; trying to understand why the wrong researchers were making the choices they made; because of 5261, spontaneous

answers came to mind, because of practice; many of the questions might be confusing for those who don't have practice answering those questions, e.g., NHST Q1;

Interview 02
- Got confused on first page; didn't notice that there weren't any questions on that page; kept trying to process the text
- NHST Q1: MISSED intended idea? Misread question?
- NHST Q2: confused; had difficulty differentiating between NHST Q1 and Q2; MISSED intended idea? 'Q1/Q2 may be a problem
- NHST Q3: Wrong answer, but can't tell if it's due to MISSING the intended idea or something else;
- NHST Q4: Not sure of their reasoning; seems off, even though they agreed with the correct answer
- NHST Q5: Wrong answer but INTENDED idea?
- NHST Q6: missed the negative in the average of trials in Team 3; more deliberation than on other questions; Team 3, because closer to 0 and other intangibles of being "balanced"
- NHST Q7: Part way on INTENDED idea? Faster reaction than on other questions
- NHST Q8: Generally correct idea, but part way on INTENDED idea? Faster reaction than other questions; mixed up sample and population;
- NHST Q9: Questioning the idea of a faulty device with the observed data; correct answer; considered the OTHER question as part of this problem; i.e., they took Q7, Q8, and Q9 into account all at once in answering Q9; ('Proposed change: somehow "delink" the implications of each concern in each of Q7, Q8, and Q9')
- NHST 10: Correct answer, mostly INTENDED idea;

- Situation: 6 minutes and 35 seconds, instead of 6.35 minutes;
- Thought the plot indicated that planes were mostly EARLY (because of so many negative values)
- Est 1: more hesitation than on other questions; wrong answer, wrong idea?
- Est 2: Noticed similarly to NHST situation; had trouble differentiating between Est 1 and Est 2, as in the NHST parallel;
- Est 3: wrong answer, not quite the intended idea?
- Est 4: paused on what Friend 1 was doing with the interval; ideas of sampling variability; correct answer, but MISSED intended idea about sim-as-hypothetical
- Est 5: correct answer, used words to indicate idea of simulation, but not quite the right reasoning
- Est 6: drawing on their answer from NHST; FORGOT THE SAMPLE SIZE and was trying to refer back to it; more deliberation than on other questions;
- Est 7: Referred back to NHST section; correct idea, INTENDED idea;
- Est 8: Emphasized the sample size itself; MISSED idea; ('Proposed change: "Does running **trials** overcome the low sample size…"'
- Est 9: Referred back to NHST section; linked all concerns Est 7, Est 8, and Est 9 together; seemed to gloss over the issue itself for the bigger picture of all three concerns presented;

- Est 10: Disagreed with BOTH! Took the concerns from Est 7, Est 8, and Est 9 as TRUE, and therefore disregarded an exact replication ('Proposed change: move Est 10 and NHST 10 to be PRIOR to the three concerns'); but in general, did not show the misconception;

- Overall: would not have been able to answer the questions without taking the stats class that was taken;

Interview 03
- Est Q1: a reaction, incorrect, MISSED idea?
- Est Q2: more reasoning than NHST Q1, incorrect, MISSED idea?
('Proposed change: Est Q1 and Q2 and NHST Q1 and Q2 seem to be really slowing interviewees down')
- Est Q3: hard to follow reasoning, bringing in many unrelated ideas
- Est Q4: can't remember the RQ; ('Proposed change: Friend 1's answer didn't match participants expectation of how to answer the RQ, so they went with Friend 2; thus, Friend 1's answer should be changed so that it actually answers the research question')
- Est Q5: MISSED idea? focused on something about a margin or interval
- Est Q6: brought in a number of issues, part of answer might be toward intended idea, but hard to tell;
- Est Q7: INTENDED idea, correct answer;
- Est Q8: mostly intended idea, but focused on sample size need; focused on authority of teacher ('Proposed change: change Teacher to someone else')
- Est Q9: INTENDED idea, correct answer
- Est Q10: focused on problems with data; (linkage of previous questions to this one)

NHST intro section: had to verify what positive and negative meant after seeing the sample data
- NHST 1: recognized situation from in class, which led to answer; mixed up this simulation with the last simulation;
- NHST 2: trying to predict how they would answer the research question before seeing any further information needed to do so;
- NHST 3: focused on lack of an interval as reason;
Wants RQ on each page
- NHST 4: bringing in compatibility interval in lieu of p-value, led to wrong answer;
- NHST 5: not remembering p-value, which was not formally covered in this person's class;
- NHST 6: focused on size of compatibility interval;
- NHST 7: INTENDED idea, correct answer;
- NHST 8: focused on sample size itself and not on role of simulation, which implicitly gets at INTENDED idea;
- NHST 9: correct answer, but not necessarily addressing the role of simulation;
- NHST 10: correct answer, but bringing in Supervisor's concerns, seemed to focus on wrong idea; focused on sample size overall

Interview 04

Overall: high level of experience and correct thinking leading to longer response times and breaking of items; item suggestion: sample size (NHST9) may just need to be more direct, e.g., “Does running many trials of the simulation help increase the sample size to address this concern?”

“Diastolic” tough word;
Double checking what positive and negative values mean
NHST1
INTENDED idea; correct; how they answered this question invoked NHST2; NHST1 and 2 may be inseparable questions in some minds, indicating it might not be worth asking both;
NHST2
INTENDED, correct.
NHST3
Interesting interpretation; interesting paths of reasoning; something about this item and the suggestion of Researcher 1 may be inputting ideas into people’s brains they hadn’t thought of
NHST4
Good ideas, correct, INTENDED idea
NHST5
Focused on p-value definition;
NHST6
“More simulations, more data”; (I would like to probe more into what is meant by “data” in these thoughts; “data” seems to be generally used term.); fascinating answer; wanting to publish all four teams; BROKE ITEM PARAMETERS (‘given the knowledge and experience of this participant, that might not be reflective of the target population, but it indicates that the item; if anything it shows the bounds of interpretation’)
NHST7
Sophisticated response, INTENDED idea; expanded beyond the correct answer to a more sophisticated design (‘participant’s high level of experience seemed to break item parameters again’)
NHST8
INTENDED idea; correct
NHST9
Invoked sophisticated idea about representativeness; didn’t quite get at intended idea; had some other ideas about what simulation could be helpful fo
NHST10
BROKE ITEM PARAMETERS by picking apart “consistently higher reading” -> used that as a way to address the confound with additional work by the researchers;

Airplane situation: spent much time dissecting the skew and what it meant.

Est1
Invoked sophisticated idea, not quite intended idea; focused on sampling with replacement as mechanism, not quite on the conceptual path; tripped up on the assumption about “variability”;

Interview 05
Overall:
- First two questions in both sections created lots of difficulty;
- Having a missed focus on Est1 may disrupt to ability to answer Est2; 'difference of only 0.04 between 6.35 and 6.39 seems to cause a focus on the choice between the numbers not mattering, which is NOT the point of the question; may need to either change the numbers OR simply point out bootstrap average vs. sample average in generality; the point is to attend to the idea of sim vs. real average, not 6.35 vs. 6.39 as numbers themselves;
- At multiple points, DID correctly state assumption needed for Est1 and 2 that bootstrapping assumes the sample is a stand-in for the population

Est1
No immediate answer; not on intended idea; 'This question perhaps could be simplified. Perhaps: What is the simulation designed to model?'
Est2
Not on intended idea; focused on sampling; focused on the skew of the graph; word "assumption" lead to distraction about assumptions of linear regression
Est3
Minimal difference in 6.35 and 6.39 led to increased disregard for the need for the question. INCORRECT ANSWER WITH INTENDED IDEA (prioritized bootstrap average as having more samples)
Est4
Est5
Est6
Eventually got a correct idea (about smoothing out shape of distribution) after probing; initially focused on precision
Est7
Sample size focus
Est8
('Both this and Est10 may not be leading participants to attend to the simulation aspect; there may be a more direct way to ask this: E.g., did the bootstrap sim alleviate this concern?')
Est9
Focused on a variety of unintended issues, and then INCORRECT INTENDED IDEA emerged with probing
('This question may need to be more direct: Suppose the researchers want a larger sample size for their study. Researcher 2 suggested: Let's run more trials of the bootstrap simulation increase the sample size.' Do you agree the suggestion of Researcher 2? Why or why not?')
Est10

N1
Word "assumption" causing problems;
N2

N3
Skipping beyond question to discuss confidence interval calculation; had to be prompted to answer question as stated
N4
Thought that we were bootstrapping, which seems to have led to a different thought process

Interview 06
Assumed the graph may need to be memorized; spent extra time on it.
'Add "minutes" to 6.35 in the description of the simulation.'
'Add "minutes" to interval in E4.'
'Add "Using the formula above, Friend 1 calculated…" in E4'
They want the RQ box on each page to be shown more as a reminder, instead of appearing like a new RQ
'Consider making bootstrap average further away from the sample average; the tiny difference of 0.04 as not really mattering was noted by this participant.'
Participant appeared to confuse sample units vs. trials, particularly when interpreting graphs.

E1
NOT on intended idea; began to get closer to intended idea with probing about knowledge of bootstrapping
E2
Partially answered E2 while considering E1; vague language that is generally correct, but not quite on intended idea;
E3
ON INTENDED IDEA AND INCORRECT! Bootstrapped average is viewed as better; it seems participant is mixing up the sample graph with the trial graph
E4
Apparently on intended idea; concerned about width of interval
E5
Interesting idea: connecting how bootstrapping results are representative of sample data; INCORRECT, SOMEWHAT ON INTENDED IDEA: 6.39 touted as more accurate; tiny difference in values noted
E6
Hard to tell if on intended idea or not;
E7
Expressed difference between simulation studies vs. replication studies; correct idea, apparently on point;
E8
Correct. Not on intended idea; 'This item seems to immediately cue participants that there is a problem with the data, and the idea that the simulation may play a role in alleviating the problem does not seem to be considered. I think I would like them to at least consider this possibility, which the current phrasing doesn't seem to initiate.'
E9

On generally intended idea suggesting that this participant does not hold Panacea misconceptions; correct.
E10
Idea about quantifying the device error; 'I would like participants to focus on the role of the simulation, but that doesn't seem to be considered directly;'

Interview 07
Overall:
Wasn't sure if they needed to memorize graph or not.
'Consider moving N5 to another page, to prevent influence on N4.'
Was concerned that RQ didn't address the specific population ("males"); 'consider removing all demographics in research situation introduction and making a nationally-representative sample'
'Consider adding on the two situation pages: "You do not need to memorize this graph".'
Minor difference between 6.35 and 6.39 factored into E3 answer. 'Pick a more discrepant trial average.'

N1
Was curious about number of trials in lead up to this item.
Questioned the "correctness" of the simulation.
Not on intended idea.
N2
Vague.
N3
Wasn't sure whether to "trust" the simulation or not.
Went down a deduction path questioning why the simulation did what it did in the first place.
N4
Was hesitant to agree with either researcher, due to lack of knowledge.
N5
Suggested that N5 might have informed how they answered N4, had they scrolled down further to see N5 while answering N4.
N6
Minor issue with word "where".
Tried to bring in context knowledge about blood pressure to interpret the averages of the trials.
N7
Focused on need for not-males ('RQ didn't specify population, but study focused on males').
N8
N9
N10

Est situation:
Concerned that RQ is focused on "delays", but technically, not all flights are delayed.
'Perhaps rewording is needed'

E1
Didn't know. Not on intended idea.
E2
Somewhat closer to intended idea.
E3
Hung up on word "interval".
E4
Used information from this, E4, to consider going back to change E3. Not really on intended idea.
Since the formatting Friend 2's response more closely matched the research question format, they disregarded the interval.
E5
Concerned about multiple aspects in Friend 3's answer that could lead to someone only partially agreeing with Friend 3's assertion.
E6
Concerned about sample size differences between this section and the last section.
Fascinating misinterpretation of why different numbers of trials are appropriate for different sample sizes.
E7
Concerned about "sample adequacy".
Was confused by who "friends" were (i.e., Friend Group? The original friends?)
E8
Mixed up details of the study in answering.
E9
On intended idea!
E10
On intended idea!

Interview 08
Overall
"Diastolic" tough word
Thought the RQ on the introductory page was a question.
Participant did not seem to have experience with randomization tests, leading to difficulty in reasoning through NHST questions;
CHANGE "10 QUESTIONS" to "9 QUESTIONS" in the two section intros!
Add "Why or Why not" to N8

N1
Either not on intended idea, OR, on intended idea but incorrect.
Questioned the correctness of the procedure; 'Participant may not have experience with randomization tests; they referred to the procedure as "bootstrapping"'
N2
Indicated that they didn't have experience with "graphs"; not quite on intended idea;
N3
Forgot what a low p-value meant
N4

Was thrown by the inconsistent pattern in the p-value;
N5
On intended idea, but lack of understanding of randomization test seems to be interfering with answers;
N6
Misunderstood what replication meant in this setting; on intended idea with incorrect thinking, potentially;
N7
Brought in more sophisticated ideas about controlling for variables;
N8
No explanation, as question was missing "Why or why not?"
N9
Brought in idea about "ethics";

E1
On intended idea with apparent misconception!
E2
Prioritizing the simulation! Some misconception about bootstrapping present; potentially on intended idea;
E3
Flagged the word "certain" and that dissuaded the participant away from the wrong answer;
E4
Maybe on intended idea;
E5
Might be on intended idea;
E6
Submitted unintended answer and then spoke intended answer after clicking past the page; they DID provide the correct answer;
E7
E8
On intended idea;
E9

Overall reactions:
Noticed ideas about replication and validity;
Thought some of the information was superfluous, e.g., 2019;
Had headache;

Interview 09
Overall
Seemed to respond well to the idea of "friends" being part the characters in the Estimation situation;

Intro
Paused on 6.35 minutes (Questions 6.35 minutes vs. 6 minutes and 35 seconds)

Began taking paper notes
Looking for correspondence between research question and sample of data
Questioning whether 2019 is needed, but seems to support including it

E1
Mostly on intended idea, correct idea;
Would like to draft answer and then come back to it after learning what the other questions are;
E2
Would like to consult a textbook to answer this question; on intended idea, correct idea;
E3
After reading this question, they wanted to go back and change E2; worked back from Friend 1's interval to verify that 6.35 was used as the sample estimate;
Participant flagged "certain" in Friend 2's answer; 'consider removing "certain" from their answer'
E4
Thought 6.10 was "significantly enough lower" to do "more testing" to evaluate the difference between 6.10 and 6.35;
'Consider changing "trust" to something else in this item; e.g., "accuracy"
E5
'Consider changing to something like: Is there anything gained by using a larger number of trials? If so, why? If not, why not?'
E6
Remembered Friend 2 as being overly confident about 6.10 as answer to RQ, and thus Friend 2's replication approach surprised them; on intended idea, correct idea;

Didn't want to agree with either; didn't want to replicate the same experiment as they thought 500 trials wasn't enough

'Consider changing to: "Which of the two approaches would you prefer to use for a replication study? Why?" This prevents disagreeing with both, forcing a choice, even if the approach isn't perfect from their perspective'
E7
Seems on intended idea;
E8
Recalled (perhaps incorrectly) that bootstrapping is used for instances of a small sample size;
'Consider changing to something else; the idea of using bootstrapping for small samples might be interfering with the intended idea of trials vs. sample size; maybe something like, "There was desire to get a higher sample size for the study. Would the fact that many trials were…?" '
E9
Focused on word "consistently";

Intro
Identified small sample size;

Questioning the accuracy of the simulation; participant does not seem to recognize that this is a randomization test, leading to confusion about why the simulation was set up the way it was;

Interview 10

Overall

Needed some coaching to overcome anxiety on answering questions; was told to just focus on reactions to questions and not on trying to come up with full-fledged answers;
Had to take phone-call partway through that involved good news and personal health information;
Looked up information multiple times; CONSIDER RECOMMENDING TO NOT LOOK THINGS UP DURING ACTUAL INSTRUMENT DISTRIBUTION

E-Intro

Paused and cued in positive- and negative-number convention in data;
Spent time looking up the name for a right-tailed distribution, but got on a tangent about right-tailed tests;

E1

On intended idea, apparently incorrect! (potential misconception demonstrated); thought this was a hypothesis test;

E2

Incorrect answer; potential misconception; said to take data from bootstrap sample, because it was randomized and there were more trials; 'apparent misunderstand

E3

Questioned which average was used for Friend 1's calculation, which seemed to add difficulty to answering item; since they thought 6.10 was the correct sample estimate to choose, they questioned the interval from Friend 1, as Friend 1 used 6.35.
ADD RQ TO E3 PAGE!

E4

Not on RWC, but good thoughts provided

E5

Thought more trials better; looked up information about bootstrapping to answer questions; SCRAP THIS WORDING, NOT ELICITING WHAT I WANT

E6

On intended idea, correct;

E7

Questioned whether bootstrapping might help, but ultimately landed on correct answer;

E8

On intended idea, initially correct, but questioned what sample size meant; searched for question/answer on internet directly; ultimately seemed to have correct idea;

E9

Wrong answer initially, AND THE INTENDED MISCONCEPTION!! (An interaction of the concepts of bootstrapping and confidence intervals.), but then got on an idea about adjusting the CI;

Interview 11

Overall
Reminder question: what class did you take? 5261?

Est Intro
Took note of negative and positive nature of values
E1
Some idea about randomizing things
E2
Questioned whether the early departures were included in the bootstrap trials graph
Noted that they didn't remember this;
Wrong answer, INTENDED MISCONCEPTION, but may not have articulated that in their answer
E3
Correct answer; interviewer added extra question; explanation hard to follow;
E4
Correctly pointed out that we don't know what the actual average is; correct answer;
Consider changing E4 by adding "…will tend to be more accurate than the sample average", to this question, as it's currently framed as an absolute.
E5
This new prompt elicited a much richer response than other versions of this prompt; both correct and incorrect ideas mixed together; generally, a correct answer;
E6
Somewhat agreed with Friend 2 but didn't prefer it. Therefore, they held some degree of misconception, but still chose the correct answer;
Interviewer extra question about whether or not there is any validity to Friend 2;
Need to change Friend 2's response: maybe something like "Is Friend 2's approach a valid way to consider a replication study?"
E7
Correct answer
E8
Somewhat eliciting the misconception! Hard to tell what the conception of "sample" is in their hand;
E9
Correct idea, answer;

N situation
Slightly confused by positive and negative values
N1
Somewhat on correct idea, though they mainly restated the first sentence of the explanation; may need to eliminate the "random chance" explanation in instructions.
N2
Questioned what was done in the simulation, perhaps suggesting
Doesn't appear to understand the purpose of the randomization simulation, which led to questioning, confusion; though, they somewhat communicated correct ideas about randomization; a bit of a hodge-podge;
N3

They questioned the fact why only trials larger than a value were taken; led to essentially a non-answer to this question; (they were likely thinking of a two-tail test and were thrown by a one-tail)
N4
N5
Took note of all results centering around 0; This new phrasing of the item seemed to work again; ALTHOUGH, they indicated misconception about centering on 0; then, wanted to go back and double-check the introduction and simulation;
Is there a way to have the simulation available to them?
N6
INTENDED MISCONCEPTION
N7
Correct idea
N8
Correct idea; apparently clearer answer here than on E8; some study design misconception shown
N9
Correct idea;

Interview 12
Overall
Didn't type answers to all questions, though did speak answers to all questions;
Felt like they forgot a lot;
Was questioning whether a back-button was needed; resolved by the RQ being at the top of the page;

N_intro
Went back to verify positive and negative value meaning
N1
Might be on intended idea; focused on bias, which may indicate real-world misconception;
N2
Focused on shape of distribution; correct answer, but not necessarily on correct idea;
N3
Introduced some incorrect idea about repeatedly running the simulation and sample size; incorrect thinking
N4
N5
Focused on reduction of p-value for more trials AND focused on shape of distributions; generally on correct idea
N6
Right answer with wrong argument; focused on removing bias with a new sample;
N7
Confused on the role of randomly assigning values, indicating not understanding the original purpose of the randomization; still, correct answer;
N8

Instinctively correct answer, but they couldn't explain why;
N9
Correct idea; brought in sophisticated thoughts about accounting for the consistency in the measurement flaw;
E_intro

E1
Not quite on correct idea;
E2
Correct answer at first, then silent, then chose incorrect answer, focusing on comparing things to the interval;
E3
Incorrect answer;
E4
Completely unsure at first; wasn't sure to the extent that differing values was a problem;
E5
Generally, on correct idea;
E6
Same general answer as N6
E7
Correct idea;
E8
No idea for an explanation
E9
Same answer as N9

Appendix C2: Instrument changes throughout interviews

Changes to the 1st generation of the instrument

6/25/2020

After three interviews, the following changes were implemented

- Reordering items:
    - N07 -> N08
    - N08 -> N09
    - N09 -> N10
    - N10 -> N07
    - Est07 -> Est08
    - Est08 -> Est09
    - Est09 -> Est10
    - Est10 -> Est07
    - Reason: to prevent the idea of data problems from biasing participants on thinking about a replication study; some participants thought the data were invalid from earlier items, and thus, the replication study was dismissed; moving the replication item before the hypothetical data concerns should prevent this dismissal
- Adding page break after each old N07-N09 and Est07-Est09
    - Each of the above items on their own page
    - Reason: to limit the amount of inter-item conceptual linking and to reduce the number of words on a page; this should help separate the ideas of each item, by containing them to a single page
- Rewriting old N07-N09 and Est07-Est09 to NOT have an authoritative actor
    - "Supervisor" and "Teacher" removed
    - Reason: Undue authority may be given to these actors, causing the participant to disregard the reasoning from the subordinate, as demonstrated by one participant
- Adding "Research Question" to the top of each page on Qualtrics
    - Reason: participants wanted to refer back to the research question, which was mostly only presented on the first page of each section
- Rewriting the focus of old N08 and Est08
    - Changed text in the prompt to explicitly ask participants to address simulation trials vs. sample size
    - Reason: all participants seemed distracted by or fixated on the idea of a "small sample size", which appeared to lead them away from discussing the intended topic of the question
- Rewriting old N07-N09 and N07-N09 in general
    - Changed the text given the above issues and to simplify wording

- Reason: the above issues need addressing and other rewording should help improve flow of these items

6/29/2020

From Bob's feedback:

- Minor rewrite of Est10 to increase parallel nature with N10
  - Removed word "much" from old Est10

Following implementation of the above changes from 6/26 and 6/29, the interview instruments will be renamed as follows:

- instrument_interview_A1 > instrument_interview_A2
- instrument_interview_B1 > instrument_interview_B2
- instrument_interview_C1 > instrument_interview_C2
- instrument_interview_D1 > instrument_interview_D2

Changes to the 2nd generation of the instrument

7/8/2020

Proposed changes to consider:

- I propose eliminating n/est02:
  - From a test blueprint standpoint, the RWC numbers work out fine:
    - Without n/est02:
      - Process: 4 (2 shared with Panacea)
      - Product: 3
      - Panacea: 4 (2 shared with Process)
  - From a think-aloud standpoint, these questions seem to have caused the most trouble and participants have not quite been on the intended idea; one participant couldn't even distinguish between n01 & n02 (and est01 & est02)
- Combine n/est01 and n/est02 into one question:
  - The goal should be to focus on the RWC of PROCESS (with the PRE of Purpose)

  A general question could be this, but it's TOO vague: Consider the specific steps of the simulation above. How does running this simulation help answer the research question?

- Instead, a more-assumption based item, like the original iteration of n/est01 could be used, with slight rewrites:
  N01: Consider the steps of the simulation above. What assumption about the effect of fish oil relative to soybean oil is the simulation designed to model?

  N01: How does running the simulation above allow the researchers to say whether the effect of fish oil is different from soybean oil?

  N01: Consider the steps of the simulation above. Why is this particular simulation appropriate for answering the research question?

  N01: What does this simulation assume to be true about the effect of fish oil vs. soybean oil, such that the researchers can answer the research question?

  Est01: Consider the steps of the simulation above. What assumption about the sample and the population from which it came is the simulation designed to model?

  Est01: How does running the simulation above allow the researchers to estimate the uncertainty in the sample average of 6.35?

  Est01: Consider the steps of the simulation above. Why is this particular simulation appropriate for answering the research question?

  Est01: What does this simulation assume to be true about the sample, such that the researchers can answer the research question?

- n/est08: Suppose there was a confounding variable in the study. Suppose that all of those in the fish oil group were taking a blood pressure medication, while all of those in the soybean oil group were not.

Since each trial of the simulation randomized blood pressure values between the two groups, does the running the simulation still allow the researchers to provide a valid answer to the research question? Why or why not?

- n/est09: Suppose there was concern that the sample size was too small.

  Is running the simulation more times a valid strategy for increasing the sample size for the study? Why or why not?

- n/est10: Suppose that the medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have.

  Does running the simulation still allow the researchers to provide a valid answer to the research question? Why or why not?

7/15/2020 & 7/16/2020

Changes to be implemented in 3rd generation of instrument: see tracked-changes document of *qualtrics_rwc_draft_master_edits_2020_07_15*.

Summary:

- Added "Reminder of" to each additional page where the research question appeared, so as to not suggest a new research question was being considered on each page
- Replaced n01/n02 with the following: "What is the purpose of randomization in the simulation (in terms of how it helps you answer the research question)?"
- Replaced e01/e02 with the following: "What is the purpose of resampling in the bootstrap simulation (in terms of how it helps you answer the research question)?"
- Rewrote study text for NHST to avoid specific demographics.
    - Focused on general "population of interest", with "random sampling" to eliminate any concern about generalizability

    - Per the above change, also changed the research question to include "in the population" to for a parallel reference to a population as is done in the Estimation research question.
    - Also, removed the word "group" throughout research question to simplify wording.
- Added page break between n04 and n05, AND between e04 and e05.
    - One participant considered changing their answer to n04 after reading n05.
    - For parallel purposes and because of how e03, e04, and e05 are worded, potential influence of e05 for the prior two items could occur.
- Added "Note: You do not need to memorize this graph." before the two sample graphs, as some participants weren't sure if they would need the information later.
- Added "minutes" on lead-in text prior to bootstrapping algorithm to clarify the 6.35
- The bootstrap 500-trails average (changing from 6.39)
    - Multiple interviewees thought the minor difference in average times (6.35 from sample vs. 6.39 500 trials), so something needs to be done to remove focus on the small difference and more the fact there's any difference at all
    - Reran the bootstrap simulation with different seeds until a larger difference from 6.35 was seen:
        - New 500 trial bootstrap mean = 6.10 minutes
        - SD still = 1.85 minutes
    - Updated Est_Fig2 to show new bootstrap 500-trial average
    - Updated e03 to indicate 6.10
    - Updated numbers in e04
    - Updated the combination plot with the four friend groups to show the new simulated mean of 6.10 for the 500 trials
- Added "confidence or compatibility" to "…interval" in text preceding interval formula, as some participants weren't sure what the generic "interval" meant.
- Intra-item parallel statement phrasing on e04
    - Two participants were thrown by simply reporting the interval. They didn't think that answered the research question directly, whereas reporting the simulated mean of what used to be 6.39 was viewed as the answer.
- Item e05 rewritten:
    - Int07 thought multiple issues were present in the item.
    - Thus, the concern was simplified down, using the following wording:
        - Suppose a third friend saw the plot of 500 trials and raised a concern.

          Friend 3: *"The average of 6.10 from the 500 bootstrap trials is different from the sample average of 6.35. The fact that these two averages are different suggests we should not trust the sample average of 6.35."*

- Item n07/e07 text addition:
    - Added "with the same sample size"
    - The low sample size in n07 seems to have distracted a few participants. This text added to try to mitigate the concern about needing more participants JUST because of low sample size. Trying to emphasize what "replication" means, independent of adequate sample size.
    - The same text was added to e07 to keep things parallel.
    - e07 statement from Friend 2 also adjusted to include "random sampling" to reflect situation description and parallel phrasing in n07.

Changes to the 3rd generation of the instrument

- 7/20/2020
- Noticed two small problems with Generation 3 versions, after A3 was taken by an interview participant
    - Test and Section Intros should say "9 questions" instead of "10".
    - N/Est08 should have "Why or why not?"

Changes to the 4th generation of the instrument

- 7/21/2020
- E3 Based on responses from Int08 and Int09
    - Friend 2's response in E3: "We can just use the average of the 500 bootstrap trials. We can be certain that the average flight delay for all departures from MSP with Delta Airlines in 2019 was 6.10 minutes."
        - The word "certain" seems to be flagging the response for being wrong; while "certain" is technically accurate, it may be distracting from the intention of the item
        - Possible new wording: "We can use the average of the 500 trials. The average flight delay for all departures from MSP with Delta Airlines in 2019 was 6.10 minutes."
- E4 based on Int09
    - Friend 3's statement in E4: "Friend 3: "The average of 6.10 from the 500 bootstrap trials is different from the sample average of 6.35. The fact that these two averages are different suggests we should not trust the sample average of 6.35."
    - The difference in values caught Int09's attention, but they gave a vague answer involving more testing
    - Possible new wording: "The average from the simulation is different than the average from the sample. In these instances, we …This suggests the bootstrap average of 6.10 is more accurate (less biased?) than the sample average of 6.35."

    - Or: “The average of 6.10 from the 500 bootstrap trials is different from the sample average of 6.35. When these two numbers are different, the bootstrap simulation is more trustworthy than the sample average.”
- E5 based on Int09 and others
    - It’s hard to tell how to score this item: unless participants say something blatantly wrong, then vague answers that are off-topic technically aren’t wrong, but that defeats the purpose
    - Perhaps, new wording like, “What is gained by running more trials of the simulation, in terms of answering the research question?”
    - Upon further review, I plan to leave this question, as is.
- E6 from Int09:
    - Int09 didn’t agree with either, but preferred the correct answer.
    - Consider alternate wording from a preferential standpoint: “Which friend’s approach would you prefer to use? Why?”
- If E6 changes, then N6 needs to change accordingly:
    - “Which researcher’s approach would you prefer to use? Why?”
- E8 from Int09:
    - Int09 thought bootstrapping was intended to be used with a small sample size, thus, the idea of trials creating more samples was lost;
    - New possible wording: “Suppose your friends wanted a larger sample size. Would the fact that your friends ran many trials of the bootstrap simulation create a larger sample size for the study? Why or why not?”
- If E8 changes, then N8 needs to change accordingly:
    - “Suppose the researchers wanted a larger sample size. Would the fact that the researchers ran many trials of the simulation create a larger sample size for the study? Why or why not?”

- 7/22/2020
- Changes implemented:
- Page break between E2 and E3
- E3 new wording:
    - Your friends also disagreed on how to answer the research question. Friend 1: “I calculated an interval using the formula. We can be 95%

      confident that the average flight delay for all departures from MSP with

      Delta Airlines in 2019 was between 2.65 to 10.05 minutes.”

Friend 2: “I looked at the average of the 500 bootstrap trials. The average flight delay for all departures from MSP with Delta Airlines in 2019 was 6.10 minutes.”

Given only these two options, whose answer do you prefer? Why?

- E4 new wording:
  - VERSIONS A/B/C/D:
    Suppose a third friend saw the plot of 500 trials and raised a concern.

    Friend 3: “The average of 6.10 from the bootstrap trials is different than the sample average of 6.35. When these two averages are different, the bootstrap average will be more accurate than the sample average.”

    Do you agree with Friend 3? Why or why not?

  - VERSION E:
    Suppose a third friend, after examining the plot of 500 bootstrap trials said the following.

    Friend 3: “The average of 6.10 from the 500 bootstrap trials is different from the sample average of 6.35. Since these two averages are different you must have done something wrong.”

    Do you agree with Friend 3? Why or why not?

- E6 new wording:
  - Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results. However, they disagreed on what to do.
    Friend 1: “We should run our bootstrap simulation again with the data we already have.”

    Friend 2: “We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample.”

    Whose approach do you prefer? Why?

- Parallel wording for N6:
    - Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results. However, they disagreed on what to do.
      Researcher 1: "We should randomly sample another 14 participants and put them through the same study that we just did."
      Researcher 2: "We should run our randomization simulation again with the data we already have."
      Whose approach do you prefer? Why?
- E8 new wording:
    - "Suppose there was concern that the size of the original sample was too small. Did the fact that your friends ran many trials of the bootstrap simulation create a larger sample size for the study? Why or why not?"
- Parallel change to N8:
    - "Suppose there was concern that the size of the original sample was too small. Did the fact that the researchers ran many trials of the simulation create a larger sample size for the study? Why or why not?"

Changes to the 5th generation of the instrument

Changes to consider:

7/28/2020

Overall, I think a version F (so, F6) could be beneficial where there's a totally different version of the n/e_trials item:

- Wording A approximately asks, "Why might changing the number of trials be related to any of the differences shown above?"
- Wording B approximately asks, "Whose group would you choose, and why is there number trials appropriate?"
- As it stands, there is NO clear correct answer to either version.
- For Wording A, there are incorrect ideas that have been communicated (e.g., see Int03). Although, vague ideas have also been stated. These vague ideas may not technically be wrong, but they aren't necessarily the point of the item. Thus, is the point to score "correct" for items that are "on point" or just "not technically wrong"? That doesn't seem like an effective way to score an item.
- Similar difficulties exist for Wording B. See Int07 for an example of not technically being wrong, but not really being on point. This person's answer

would be hard to score. What would getting "points" for this item mean? What is being measured?
- With Wording B, an idea that was stated multiple times was that there is a limit for "too many" trials, or that you can run a simulation too many times.
- Another consideration is that in EPSY 3264 [CATALST], typical trial numbers are in the hundreds, while in EPSY 5261 [Lock 5], THE typical trial number is 10,000.

One way around n/e_trials could be rephrasing as:

- Is there something being gained by the friend groups running more trials of the simulation? If so, what is it? If not, why not?
- Is there something being gained for the friend groups that more trials of the simulation than the others? If so, what is it? If not, why not?
- What is happening by running more trials of the simulation?
- Does running more and more trials of the simulation help answer the research question? Why or why not?
- Is running more trials of the simulation helping to answer…
- What is the minimum number of trials that seem necessary for answering the research question? Why this number and not something larger or smaller?
- Based on the figure, what is the minimum number of trials that should be run, in order to answer the research question?
- Is there anything that would be gained by running more than 5000 trials of the simulation? If so, what is it? If not, why not?

I think the following is the version of n/e_trials I want to try:

- Consider the results from the smallest number of trials, 100. Was there something gained by the other friend groups running more trials of the simulation? If so, what did they gain and why did more trials provide it? If nothing was gained, why not?
  (This might be easier to score than the other versions.)

Another one I like, much simpler:

- Is there any reason to run more trials beyond 5000? Why or why not?
  (This item might get at the same concept as the item above with an easier

  cognitive load to process, BUT, it might be harder to score.)

8/3/2020

- For est_interpret, Int10 questioned which average was used by Friend 1. Given the numerical confusion, I propose to remove the numbers completely, and just focus on the issue of "interval vs. center":
    - E.g., To answer the research question, Friend 1 began calculating an interval for the sample estimate. However, Friend 2 stopped them and said, "You don't need an interval to answer the research question. You can answer the research question by reporting the average of the 500 bootstrap trials." Do you agree with Friend 2? Why or why not?
- Make sure to add the RQ at the top of the page with est_interpret
- For est_replication, Int10 circled around each answer, potentially showing the misconception; but, they ultimately chose the correct answer, despite apparently holding the misconception.
    - I would like a way to rewrite n/est_replication without biasing participants away from the wrong answer and minimizing the chance they choose the wrong answer if they have the misconception
    - Potential new wording: "Since this was only one sample… Friend 1 suggested that they could rerun the bootstrap simulation Is this a valid approach for conducting a replication study? Why or why not?

Changes to be implemented:

The following is being implemented on C5 in order to create version F6 for participant Int11

- Added research question reminder on page with est_interpret
- Changed est_interpret (current est_03)
    - "To answer the research question, Friend 1 wanted to report the interval they calculated. However, Friend 2 said, "You don't need an interval to answer the research question. You can report the average of the 500 bootstrap trials."

      Do you agree with Friend 2? Why or why not?"

- Change n/est_trials (current n/est_05)
    - Est_05: "Consider the results from the smallest number of trials (100). Was there something gained by the other groups running more trials? If so, what did they gain and why? If nothing was gained, why not?"
    - N_05: "Consider the results from the smallest number of trials (100). Was there something gained by the other teams running more trials? If so, what did they gain and why? If nothing was gained, why not?"
- Change n/est_replication (current n/est_06)
    - This reflects the original version of the replication item, but with fewer words and more focused on whether or not they agree with the wrong

answer (as opposed to choosing between answers, in case they don't agree with either).

- Note that the Researcher and Friend ORDER IS THE SAME! (In all other questions, the order of answers for Researchers is flipped compared to the order of answers for the Friends.) Since this item as written does not flow as well if you reverse the order in which the statements appear, I'm opting to keep the order same. (It would also be distracting to put Friend 2 first or Researcher 2 first.)
- N_06: "Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results.

  Researcher 1: *"We should randomly sample another 14 participants and put them through the same study that we just did."*
  Researcher 2: *"We can also run our simulation again with the data we already have, as that is as good as a replication of the study."*
  Do you agree with Researcher 2? Why or why not?"
- E_06: Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results.
  Friend 1: "We should randomly sample another 150 departures and

  analyze them the same way we analyzed the original sample."

  Friend 2: "We can also run our bootstrap simulation again with the data

  we already have, as that is as good as a replication of the study."

Changes to the 6th generation of the instrument

Changes to consider:

- E4:
  - Change to "…will tend to be more accurate than the sample average.", or "…is more likely to be accurate than the sample average."
  - The absolute nature of the 6th Generation framing ("…will be more accurate…") was flagged by Int11
- E6:
  - Int11 agreed with the misconception, BUT didn't totally bite on that answer, because they preferred the correct answer
  - Consider changing to, "Is the approach by Friend 2 a valid way to do a replication study? Why or why not?"

- N1:
    - Int11 repeated “random chance” from the instructions as their answer; consider eliminating “random chance” in the instructions; OR, just accept that some people may provide cookie cutter answers
        - I’m not going to change this, as this was the only instance of this occurrence, and it does signal that Int11 didn’t understand what the actual purpose of randomization was;
- NHST RQ
    - Int11 struggled with why only trials larger than a given value were being counted for the p-value
    - This seems to be one-tail vs. two-tailed confusion
    - Indeed, the p-value is calculated as one-tailed, but the hypothesis is two-tailed.
    - Consider changing RQ to be one-tailed
    - E.g., “Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?”

8/5/2020 Changes implemented:

- Change NHST RQ to, “Is there a larger reduction in average blood pressure for fish oil compared to soybean oil in the population?”
- Change E4 (Est_center) to, “…the bootstrap average is more likely to be accurate than the sample average.”
- Change E6 (Est_replication) to, “… Is the approach by Friend 2 a valid way to do a replication study?”
- Parallel change to N6 to, “… Is the approach by Researcher 2 a valid way to do a replication study?”

Appendix C3: Item history

NHST (n)

(item removed) (n) Assumption
- Generation 1-2

What assumption about the effect of fish oil relative to soybean oil is the computer simulation designed to model?

- Generation 3-7

n/a

Discussion: See n_purpose below.

(item removed) (n) RQ
- Generation 1-2

How does the assumption you just discussed help answer the research question of whether there is a difference in the average blood pressure reduction between the two groups?

- Generation 3-7

n/a

Discussion: See n_purpose below.

1. (n) Purpose
- Generation 1-2

n/a

- Generation 3-7

What is the purpose of randomization in the simulation (in terms of how it helps you answer the research question)?

Discussion: n_assumption and n_rq were eliciting unintended or vague responses. Some participants couldn't distinguish between these two items. Since the original goal was to elicit ideas about the purpose of simulation, these two items were combined into

n_purpose. Responses were either more on the point of the purpose of simulation or easier to score as being wrong or not quite on point, relative to n_assumption and n_rq.

2. (n) Center

- Generation 1-7

After seeing the distribution of 500 trials, Researcher 1 offered this observation:
"Since the results center around 0, that suggests there is no difference in effect between fish oil and soybean oil in reality."
Do you agree with Researcher 1? Why or why not?

Discussion: Item elicited responses that were on-point, vague, or off-point due to participant seemingly not possessing sufficient knowledge about a randomization simulation to have a clear answer to this item. This was apparent from several participants who disagreed with Researcher 1 (correct response) but gave vague off-point answers (incorrect or unclear reasoning). Such responses would likely either be scored as "0" or partial credit. For the participant who seemed to know the most about the instrument's content, overall, their response was correct and on point. Since this item directly asks about the intended misconception, as worded, rewording to try to elicit more on-point incorrect answers does not seem worth the extra time and investment for this instrument. Given that Phase II participants will have just seen the relevant course information for this item, the issue of lacking sufficient awareness about a randomization simulation is more likely to be addressed, such that there will be fewer off-point, vague responses.

3. (n) Calculate

- Generation 1-7

The researchers agreed to calculate a p-value but disagreed on the procedure.
Researcher 1: "Start at the average of the 500 trials, and then calculate the proportion of all trials larger than that."
Researcher 2: "Start at the sample average difference of 7.7, and then calculate the proportion of all trials larger than that."
Which researcher do you agree with? Why?
Discussion: Two ideal correct responses were elicited, as well as other incorrect responses that seemed fairly clear in the misconception that the participant had. No wording or other problems with this item observed based on responses. Predicted mild difficulty in scoring some responses with incorrect reasoning.

4. (n) Interpret

- Generation 1-7

Researcher 2 calculated a p-value of 0.014. However, Researcher 2 was unsure of how to answer the research question.
Possible answer 1: “The low p-value indicates that an average difference of 0 is not supported. There is evidence for a difference between the groups.”
Possible answer 2: “The low p-value indicates that the sample average difference of 7.7 from the data is not supported. There is no evidence for a difference between the groups.”
Which answer do you agree with? Why?

Discussion: Correct responses were a mixture of clear, ideal explanation and cookie-cutter p-value interpretations. The only incorrect response was from a participant who had forgotten how to interpret p-values.

5. (n) Trials
- Generation 1-5

Consider the differences in the four sets of results above. Why might changing the number of trials be related to any of these differences?
OR
If you had to choose the results from one team to use, which team would you choose? For the team you chose, why is their number of trials appropriate for answering the research question?

- Generation 6-7

Consider the results from the smallest number of trials (100). Was there something gained by the other teams running more trials? If so, what did they gain and why? If nothing was gained, why not?

Discussion: Responses are all over the map. For the “choose” version, two answers are clearly on point, but in general, this might be a hard item to score. The issue is that there are multiple correct selections, thus scoring for conceptual accuracy requires a clear explanation. The type of ideal explanation might be too difficult for the target population. One participant broke the item, as well, by choosing all four sets of results. Responses to the “differences” version were off point, both probably scored as “0”. I think the “gain” version is more directly worded toward what I want to elicit. The two responses were mixed; one was on point and one was non-specific. I contend it might be worth choosing the “choose” version, because it did elicit some good responses. Versions of this item were also influenced to responses to the parallel estimation item.

6. (n) Replicate (in last position of this section only for Generation 1)

- Generation 1-2

Since this was only one study the researchers wanted to do a replication study, to verify the results they just found. However, they disagreed on what to do.
Researcher 1: “We should recruit 14 new participants and put them through the same study that we just did.”
Researcher 2: “Collecting new participants is unnecessary. We can just run our simulation again, as that is as good as a replication study.”
Which researcher do you agree with? Why?

- Generation 3-4

Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results. However, they disagreed on what to do.
Researcher 1: “We should randomly sample 14 new participants and put them through the same study that we just did.”
Researcher 2: “Collecting new participants is unnecessary. We can just run our simulation again, as that is as good as a replication study.”
Which researcher do you agree with? Why?

- Generation 5

Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results. However, they disagreed on what to do.
Researcher 1: “We should randomly sample another 14 participants and put them through the same study that we just did.”
Researcher 2: “We should run our randomization simulation again with the data we already have.”
Whose approach do you prefer? Why?

- Generation 6

Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results.
Researcher 1: “We should randomly sample another 14 participants and put them through the same study that we just did.”
Researcher 2: “We can also run our simulation again with the data we already have, as that is as good as a replication of the study."
Do you agree with Researcher 2? Why or why not?

- Generation 7

Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results.
Researcher 1: “We should randomly sample another 14 participants and put them through the same study that we just did.”

Researcher 2: “We can also run our simulation again with the data we already have, as that is as good as a replication of the study."
Is the approach by Researcher 2 a valid way to do a replication study? Why or why not?

Discussion: Final version of item most directly addresses the misconception, without leaving space for unintended issues to be brought up as part of the answer. See parallel estimation item for the types of issues prior wordings were unhelpfully eliciting. The goal of this item is to address whether or not re-running a simulation is a valid form of replication. When choosing between two approaches, this issue may get lost in the situation. The one response to Generation 7 was fairly on point. Additionally, participants may not agree with either option, which occurred on least one item that required choice between two options. Forcing agreement with a statement, as opposed to two approaches more directly targets the issue, while still allowing participant disagreement with the approaches.

Changes to and maintain parallel phrasing with the equivalent estimation item also influenced this.

7. (n) Confound
- Generation 1

Supervisor: “All of those in the fish oil group were taking a blood pressure medication, while none of those in the soybean oil group were taking any blood pressure medication.”
Researcher 1: “Since each trial of the simulation randomly assigned the blood pressure values to the two groups, the presence of the medication is not a problem.”
Do you agree with Researcher 1? Why or why not?

- Generation 2-7

Suppose that all of those in the fish oil group were taking a blood pressure medication, while all of those in the soybean oil group were not.
Since each trial of the simulation randomly assigned the blood pressure values to the two groups, could the researchers still provide a valid answer to the research question? Why or why not?

Discussion: Once the “supervisor” character was removed, responses were on point and mostly clear, correct. No apparent issues with this item.

8. (n) Sample Size
- Generation 1

Supervisor: “The sample size is only 14. This seems too small.”

Researcher 2: “The small sample size is OK, because we ran many trials of the simulation.”
Do you agree with Researcher 2? Why or why not?

- Generation 2-3

Suppose there was concern that the sample size was too small.
Would the fact that the researchers ran many trials of the simulation address this concern about sample size?

- Generation 4-7

Suppose there was concern that the size of the original sample was too small.
Did the fact that the researchers ran many trials of the simulation create a larger sample size for the study? Why or why not?

Discussion: The “supervisor” character created problems in Generation 1. The Generation 5-7 version is the most direct approach for the point of the item: whether or not trials are related to sample size. Responses to Generation 5-7 were mixed, limited. Responses to Generation 2-4 were mostly on point, but two of the three overemphasized external validity and sample size. See parallel estimation item for more discussion.

9. (n) Device
- Generation 1

Supervisor: “The medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have.”
Given the simulation that was run, can the researchers still provide a valid answer to the research question? Why or why not?

- Generation 2-7

Suppose that the medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have.
Given the simulation that was run, could the researchers still provide a valid answer to the research question? Why or why not?

Discussion: Once “supervisor” character was removed in Generation 1, responses to Generation 2-7 were reasonably on point and easy to score. One response invoked a sophisticated idea of accounting for the device difference by recalibrating the data, provided the effect of the device was consistent. No major problems expected with scoring.

ESTIMATION (e)

(item removed) (e) Assumption
- Generation 1-2

What assumption about the sample and its variability is the computer simulation designed to model?

- Generation 3-7

n/a

Discussion: There were no completely correct responses. Many good ideas were offered, but each slightly to moderately missed the point of the item. The word “assumption” threw off some participants. This item was removed in favor of combining with e_rq below, to make the e_purpose item to more directly address the purpose of the bootstrap simulation.

(item removed) (e) RQ
- Generation 1-2

How does the assumption you just discussed help answer the research question about the average delay time for all Delta Airlines flights leaving MSP airport in 2019?

- Generation 3-7

n/a

Discussion: One participant had trouble distinguishing this item from e_assumption above; they simply repeated their answer from e_assumption. No correct responses were given. A few real-world conflation misconceptions were seemingly apparent. In general, responses were not discussing what I was hoping they would. Given the struggles with e_assumption and similar item issues with n_assumption/n_rq, targeting the idea of this item seemed it would benefit from being rewritten into e_purpose below, with a more direct question about purpose.

1. (e) Purpose (though mislabeled as “e assumption” in response spreadsheets for int08 through int12)
- Generation 1-2

n/a

- Generation 3-7

What is the purpose of resampling in the bootstrap simulation (in terms of how it helps you answer the research question)?

Discussion: This item has two correct types of responses (mimicking resampling from the population, OR estimating variability). Only one of five responses was correct, and it addressed variability estimation. Other responses likely all would have been scored as '0', and the types of ideas brought up easier to score as '0' for this item, relative to e_assumption and e_rq. One real-world conflation misconception also seems to have been stated. Given the difficulty of eliciting a specific set of purposes for bootstrapping, this item seems fine for an open-ended initial instrument.

2. (e) Calculate

- Generation 1-2

Your friends agreed to use 1.85 for the Standard Deviation of trials but disagreed on the Sample Estimate.
Friend 1: "The sample average of 6.35 minutes should be used."
Friend 2: "The average of the 500 bootstrap trials, 6.39 minutes, should be used."
Which friend do you agree with? Why?

- Generation 3-7

Your friends agreed to use 1.85 for the Standard Deviation of trials but disagreed on the Sample Estimate.
Friend 1: "The sample average of 6.35 minutes should be used."
Friend 2: "The average of the 500 bootstrap trials, 6.10 minutes, should be used."
Which friend do you agree with? Why?

Discussion: The bootstrap simulation average was changed from 6.35 to 6.10 minutes, because
some participants were noticing that the difference between 6.39 and 6.35 was too small, practically, to warrant any concern, throughout the estimation section (including item e_center). That distracted from the point of the item, which is to identify which average the participant thinks holds more "value", to put it inexactly. Following the change to 6.10 minutes, only one response was correct with no clear explanation. However, two of the wrong responses exhibited the real-world conflation misconception very clearly, which is a good sign that this item is eliciting what is intended. This item seems to be functioning well.

3. (e) Interpret

- Generation 1-2

Friend 1 calculated an interval of [2.65, 10.05]. However, Friend 2 argued that the interval was not needed to answer the research question.
Friend 2: “We can just use the average of the 500 bootstrap trials. We can be certain that the average flight delay for all departures from MSP with Delta Airlines in 2019 was 6.39 minutes.”
Do you agree with Friend 2? Why or why not?

- Generation 3-4

Friend 1 calculated an interval of [2.65, 10.05] and answered the research question, as follows:
"We can be 95% confident that the average flight delay for all departures from MSP with Delta Airlines in 2019 was between 2.65 and 10.05 minutes."
However, Friend 2 argued that the interval was not needed, stating:
“We can just use the average of the 500 bootstrap trials. We can be certain that the average flight delay for all departures from MSP with Delta Airlines in 2019 was 6.10 minutes.”
Do you agree with Friend 2? Why or why not?

- Generation 5

Your friends also disagreed on how to answer the research question.
Friend 1: "I looked at the interval I calculated. We can be 95% confident that the average flight delay for all departures from MSP with Delta Airlines in 2019 was between 2.65 and 10.05 minutes."
Friend 2: "I looked at the average of the 500 bootstrap trials. The average flight delay for all departures from MSP with Delta Airlines in 2019 was 6.10 minutes."
Whose answer do you prefer? Why?

- Generation 6-7

To answer the research question, Friend 1 wanted to report the interval they calculated. However, Friend 2 said, "You don't need an interval to answer the research question. You can report the average of the 500 bootstrap trials."
Do you agree with Friend 2? Why or why not?

Discussion: In Generation 1-2, Friend 1’s answer wasn’t explicitly framed in terms of answering the research question while Friend 2’s framing was. This led two participants to choose Friend 2, simply because the research question was directly answered in the item. Thus, Generations 3 and later needed more parallelism between friend answers. Responses to Generations 3-4 were OK. Generation 5 was attempted to simplify the item down to choosing between responses and to eliminate confusion from different sample estimate possibilities from e_calculate. For int10, for instance, they chose the wrong

value for the sample estimate (6.10), but knew that the interval from Friend 1 in e_interpret used (6.35). They did correctly choose Friend 1 with the confidence interval but hesitated, due to thinking Friend 1 used the wrong sample estimate. For Generation 6-7, the two responses were mixed. One response potentially exhibited a real-world conflation misconception in their interview text but not their typed response. The other response correctly chose the interval, but their explanation was conceptually inaccurate regarding statistics, generally. Since Generation 6-7 frames the issue most directly and avoids concern about what the participant chose in the prior item, e_calculate, I think this item is fine.

4. (e) Center

- Generation 1-2

Suppose a third friend saw the plot of 500 trials and raised a concern.
Friend 3: “The average of 6.39 from the 500 bootstrap trials is different from the sample average of 6.35. This indicates we should not trust the sample data nor the sample average of 6.35.”
Do you agree with Friend 3? Why or why not?

- Generation 3-4

Suppose a third friend saw the plot of 500 trials and raised a concern.
Friend 3: “The average of 6.10 from the 500 bootstrap trials is different from the sample average of 6.35. The fact that these two averages are different suggests we should not trust the sample average of 6.35.”
Do you agree with Friend 3? Why or why not?

- Generation 5-6

Suppose a third friend saw the plot of 500 trials and raised a concern.
Friend 3: “The average of 6.10 from the bootstrap trials is different than the sample average of 6.35. When these two averages are different, the bootstrap average will be more accurate than the sample average.”
Do you agree with Friend 3? Why or why not?
OR
Suppose a third friend, after examining the plot of 500 bootstrap trials said the following.
Friend 3: "The average of 6.10 from the 500 bootstrap trials is different from the sample average of 6.35. Since these two averages are different you must have done something wrong."
Do you agree with Friend 3? Why or why not?

- Generation 7

Suppose a third friend saw the plot of 500 trials and raised a concern.

Friend 3: “The average of 6.10 from the bootstrap trials is different than the sample average of 6.35. When these two averages are different, the bootstrap average is more likely to be accurate than the sample average.”
Do you agree with Friend 3? Why or why not?

Discussion: As discussed in e_calculate above, the bootstrap simulation average was changed from 6.39 to 6.10 from Generation 1-2 to Generation 3-4, due to the original difference from the sample estimate of 6.35 being so small. Generation 1-2 responses contained evidence of some real-world conflation, as well as falsely mixing in other statistics concepts. No correct explanations were given. A response in Generation 3-4 suggested that “trust” was too strong a word: the participant said the sample average could still be trusted but that further testing was warranted. This indicated some misconception that might be better addressed with different item wording. There was one response from Generation 5 that used the “you must have done something wrong” version. The participant didn’t think the difference was notable. Since the purpose of the item is to address discarding the sample, the idea of “accuracy” was invoked for the other version of Generation 6 and of the only version of Generation 7. The only participant for Generation 6 offered a partial-credit response that wasn’t clearly on point. The only participant for Generation 7 admitted to having no idea.

My proposal for this item is to invoke trust, but to do so in a more direct, comparative manner:

“Suppose a third friend saw the plot of 500 trials and raised a concern.
Friend 3: “The average of 6.10 from the 500 bootstrap trials is different from the sample average of 6.35. When these two averages are different, this suggests the bootstrap average is more trustworthy than the original sample average.”
Do you agree with Friend 3? Why or why not?”

5. (e) Trials
- Generation 1-5

Consider the differences in the four sets of results above. Why might changing the number of trials be related to any of these differences?
OR
If you had to choose the results from one friend group to use, which friend group would you choose? For the friend group you chose, why is their number of trials appropriate for answering the research question?

- Generation 6-7

Consider the results from the smallest number of trials (100). Was there something gained by the other groups running more trials? If so, what did they gain and why? If nothing was gained, why not?

Discussion: Versions of this item were influenced by changes and responses to the parallel NHST version. “Choose” responses were mixed in terms of some being hard to score and some being clear, accurate responses. “Differences” responses were also mixed in terms of focusing on the intended idea vs. not and being easier or more difficult to score. In Generation 6-7, both “gain” responses were on-point and elicited what I was hoping for. However, one “gain” response was not clear from the typed response but well argued in the transcript. Similarly, no typed response was given for the other “gain” response but was nicely explained in the transcript. I think the “choose” phrasing is too difficult to score and the “differences” phrasing is too open-ended. I propose to stick with the “gain” phrasing.

Another issue with the “choose” phrasing is that the different trial numbers common to each class (typically 500 trials for 3264 [CATALST] and typically 10,000 trials for 5261 [Lock 5]) may heavily influence what trial amounts participants think are sufficient. I don’t want that issue to be the primary aspect of the responses. The “gain” phrasing avoids this issue.

6. (e) Replicate (in last position of this section only for Generation 1)

- Generation 1-2

Since this was only one sample of data, your friends wanted to do a replication study to verify the results they just found. However, they disagreed on what to do.
Friend 1: “Collecting more data is unnecessary. We can just run our bootstrap simulation again, as that is as good as a replication study.”
Friend 2: “We should collect another sample of 150 departures and analyze it the same way that we analyzed the original sample.”
Which friend do you agree with? Why?

- Generation 3-4

Since this was only one sample of data, your friends wanted to do a replication study with the same sample size, to verify their results. However, they disagreed on what to do.
Friend 1: “Collecting more data is unnecessary. We can just run our bootstrap simulation again, as that is as good as a replication study.”
Friend 2: “We should randomly sample another 150 departures and analyze them the same way that we analyzed the original sample.”
Which friend do you agree with? Why?

- Generation 5

Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results. However, they disagreed on what to do.
Friend 1: “We should run our bootstrap simulation again with the data we already have.”

Friend 2: “We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample.”
Whose approach do you prefer? Why?

- Generation 6

Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results.
Friend 1: "We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample."
Friend 2: "We can also run our bootstrap simulation again with the data we already have, as that is as good as a replication of the study."
Do you agree with Friend 2? Why or why not?

- Generation 7

Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results.
Friend 1: "We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample."
Friend 2: "We can also run our bootstrap simulation again with the data we already have, as that is as good as a replication of the study."
Is the approach by Friend 2 a valid way to do a replication study? Why or why not?

Discussion: Changes to and maintaining parallelism with the equivalent NHST item influenced this item’s trajectory. Per comments on the NHST item, the final version seeks to avoid situations where participants don’t agree with either approach, explicitly as stated. The one response to Generation 6 exemplifies the need for Generation 7 phrasing: the participant suggested that rerunning the bootstrap simulation was a technically fine way to do a replication, but that they preferred the (correct) approach of gathering a new sample. This was hard to tell based on the typed response alone. Thus, asking the direct question in Generation 7 about validity seems essential. Responses to earlier Generations were mixed, with some great responses both correct and wrong though clearly exhibiting real-world conflation, as intended. Other responses were a bit tougher to consider for scoring purposes.

7. (e) Confound

- Generation 1

Teacher: “Only morning departure times (in the range of 5am – 12pm) were included in the sample.”
Given the bootstrap simulation that was run, can your friends still provide a valid answer to the research question? Why or why not?

- Generation 2-7

Suppose that only morning departure times (in the range of 5am – 12pm) were included in the sample.
Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

Discussion: As in the case with NHST items, the authority figure was removed from Generation 1 to alleviate the influence from interpreting the power dynamics and simply the intent of the item. All responses clearly correct, regardless of Generation. No item problems observed.

8. (e) Sample Size
- Generation 1

Teacher: "The sample size is only 150. This seems too small."
Friend 1: "The small sample size is OK, because we ran many trials of the bootstrap simulation."
Do you agree with Friend 1? Why or why not?

- Generation 2-3

Suppose there was concern that the sample size was too small.
Would the fact that your friends ran many trials of the bootstrap simulation address this concern about sample size?

- Generation 4-7

Suppose there was concern that the size of the original sample was too small.
Did the fact that your friends ran many trials of the bootstrap simulation create a larger sample size for the study? Why or why not?

Discussion: As in the case with NHST items, the authority figure was removed from Generation 1 to alleviate the influence from interpreting the power dynamics and simply the intent of the item. As discussed in parallel NHST item above, Generation 5-7 was created to directly ask about the issue of more trials creating a larger sample size. Responses to Generation 2-4 were mixed. some were good and easy to score, while others were harder to score or unclear. One response was unclear from the typed response alone, but in the transcript, they did indeed express real-world conflation after two prompts by the interviewer. Thus, the intended misconception to be unearthed was there, but the item didn't elicit it. This is major rationale for Generation 5-7, which is more direct. Responses to Generation 5-7 were mixed, with one great response that would be easy to score and two responses that were less clear. I still propose sticking with the

phrasing from Generation 5-7, due to its directness relative to the more open-ended nature of Generation 2-4.

9. (e) Device

- Generation 1

Teacher: "The flight tracker at MSP airport was found to be faulty for Delta airlines. It consistently recorded departure times as being much later than it should have."
Given the bootstrap simulation that was run, can your friends still provide a valid answer to the research question? Why or why not?

- Generation 2-7

Suppose that the flight tracker at MSP airport was found to be faulty. It consistently recorded departure times as being later than it should have.
Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

Discussion: As in the case with NHST items, the authority figure was removed from Generation 1 to alleviate the influence from interpreting the power dynamics and simply the intent of the item. Most responses for Generation 2-7 on point and accurate. One participant mistakenly thought the confidence interval would take this type of error into account, but they also seemed to hedge by suggesting that the original sample should just be retaken anyway. No concerns about this item.

Appendix D: Scoring Rubrics

Appendix D1: Scoring rubric draft from validation

This document contains the 18 items from the instrument and the guidance on how to score each of those items. Images and non-item text from the instrument are not included. Each item's scoring guide describes the content of the response that must be present for the response to be scored as a 1. Example exemplary responses were written by Jonathan Brown. This version incorporates changes made during the rubric validation meetings on 1/5/2021 and 1/6/2021.

General scoring guidelines: A single response is scored as 1 or 0. A score of 1 is marked when a response meets all stated scoring criteria for the item. Blank responses are not considered responses and thus are not scored. Only responses that do not meet criteria but still contain text are marked as 0.

Section: Null-hypothesis Significance Testing

item n01_purpose

What is the purpose of randomization in the simulation (in terms of how it helps you answer the research question)?

To score 1, must have the following:

- Something about assuming, producing, seeing, or modeling no difference or no effect

Note: Referring to the real data or observed result is not needed, as the topic of the question is the purpose of the simulation from a hypothetical vs. real-world interpretation. The role of the observed data is a separate entity in this instance.

Example exemplary response:

The purpose of randomization is to see how common the observed difference in the average blood pressure reduction between the two groups would be, if there were no true difference between these groups.

item n02_center

After seeing the distribution of 500 trials, Researcher 1 offered this observation:
"Since the results center around 0, that suggests there is no difference in effect between fish oil and soybean oil in reality."

Do you agree with Researcher 1? Why or why not?

To score 1, must have the following:

- Disagree with Research 1

AND

- Something about at least one of the following:
    - The results in the picture being simulated, hypothetical, or fake
    - The simulation being designed to have results centered around 0
    - Needing to take the observed data into account

    - The graph not providing enough information to draw that conclusion
    - Needing to compute a p-value

Example exemplary response:

I disagree with Researcher 1. This is only a distribution of simulated results. It should center around 0, because that is how the randomization was designed. The observed difference needs to be taken into account to say something about the difference in reality.

item n03_calculate

The researchers agreed to calculate a p-value but disagreed on the procedure.
Researcher 1: *"Start at the average of the 500 trials, and then calculate the proportion of all trials larger than that."*
Researcher 2: *"Start at the sample average difference of 7.7, and then calculate the proportion of all trials larger than that."*

Which researcher do you agree with? Why?

To score 1, must have the following:

- Agree with Researcher 2

AND

- Something about at least one of the following:
    - Needing the sample average, observed data, or observed result to calculate a p-value
    - The purpose of the research question or study is to use the sample average
    - Researcher 1 is only looking at the null hypothesis
    - The correct procedure for generating a p-value
    - That what Researcher 2 is doing is the correct procedure for a p-value

Note: If response agrees with Researcher 1 but gives an apparently correct explanation of p-value, mark as 0, because they could be mistakenly thinking the "observed results" from the simulation are real in the case of Researcher 1.

Example exemplary response:

I agree with Researcher 2, because the p-value is the probability of seeing the sample average difference or more extreme, given the null hypothesis of no-difference between the groups.

item n04_interpret

Researcher 2 calculated a p-value of 0.014. However, Researcher 2 was unsure of how to answer the research question.
Possible answer 1: *"The low p-value indicates that an average difference of 0 is not supported. There <u>is</u> evidence for a difference between the groups."*
Possible answer 2: *"The low p-value indicates that the sample average difference of 7.7 from the data is not supported. There is <u>no</u> evidence for a difference between the groups."*

Which answer do you agree with? Why?

To score 1, must have the following:

- Agree with Possible answer 1

AND

- Something about at least one of the following:
    - The model or assumption of no difference not being supported
    - The observed difference being unlikely to have occurred by chance
    - The null hypothesis being rejected, as long as the null hypothesis is explained or indicated to be the no-difference model or assumption

---

Example exemplary response:

I agree with Possible answer 1, because the low p-value means the observed group difference is unlikely to have occurred, were there no group difference in the population. This suggests evidence for a group difference.

item n05_trials

Consider the results from the smallest number of trials (100). Was there something gained by the other teams running more trials? If so, what did they gain and why? If nothing was gained, why not?

To score 1, must have the following:

- Something about at least one of the following:
    - More trials showing or providing more possibilities, hypothetical samples, hypothetical outcomes, or randomized data [Rubric comment: This is an emphasis on the process of the simulation, instead of the product.]
    - A better or more accurate p-value [Rubric comment: This is an emphasis on the product of the simulation, instead of the process.]

Example exemplary response:

Yes, the extra trials give a better idea of the shape of the simulated distribution. This is due to seeing more of the possible results under the no-difference model.

item n06_replicate

Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results.

Researcher 1: *"We should randomly sample another 14 participants and put them through the same study that we just did."*
Researcher 2: *"We can also run our simulation again with the data we already have, as that is as good as a replication of the study."*

Is the approach by Researcher 2 a valid way to do a replication study? Why or why not?

To score 1, must have the following:

- Disagree with Researcher 2

AND

- Something about at least one of the following:
    - Replication requiring new data

    - The results not changing by reusing the same data

Example exemplary answer:

The approach by Researcher 2 is not a valid replication study. A replication study requires new data to see if the same effect appears. Rerunning the simulation with the same data wouldn't change the results and is essentially just adding more trials to the original simulation.

item n07_confound

Suppose that all of those in the fish oil group were taking a blood pressure medication, while all of those in the soybean oil group were not.

Since each trial of the simulation randomly assigned the blood pressure values to the two groups, could the researchers still provide a valid answer to the research question? Why or why not?

To score 1, must have the following:

- State that a valid answer cannot be provided

AND

- Something about at least one of the following:
    - Confounding variables, simulation, or randomization not being able to change or correct the data
    - The data being flawed or problematic

Example exemplary answer:

The researchers cannot still provide a valid answer to the research question, because the medication acts as a confounding variable in the original data. The simulation cannot correct for the confounding variable, as the simulation only acts on the data as they are following the experiment.

item n08_samples

Suppose there was concern that the size of the original sample was too small.

Did the fact that the researchers ran many trials of the simulation create a larger sample size for the study? Why or why not?

To score 1, must have the following:

- Disagree that running more trials creates a larger sample size

AND

- Something about at least one of the following:
    - Trials being different from sample size
    - Sample size being fixed
    - Sample size being the same for each trial
    - More trials simply adding more trials to the study
    - The simulation not creating a larger sample size, generally

---

Example exemplary answer:

No. The number of trials and the sample size are not related. The sample size is the same for each trial, no matter how many trials are run.

item n09_device

Suppose that the medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have.

Given the simulation that was run, could the researchers still provide a valid answer to the research question? Why or why not?

There are two different ways to score a 1.

To score a 1, response must:

- Indicate that a valid answer cannot be provided

AND

- Something about at least one of the following:
    - The confounding nature of the device
    - The data being problematic
    - The simulation not correcting a problem with the device

Alternatively, to score a 1, response must:

- Indicate that a valid answer can be provided, if the amount that the device was off was the same amount or known each time [Rubric comment: This explanation allows the device inconsistency to be taken into account.]

Example exemplary answer:

No. The researchers cannot provide a valid answer to the research question, because the simulation cannot correct for incorrectly measured data.

Section: Estimation

item e01_purpose

What is the purpose of resampling in the bootstrap simulation (in terms of how it helps you answer the research question)?

To score 1, must have the following:

- Something about at least one of the following:
    - Mimicking a process or aspect involving a population
    - Estimating variability
    - What estimating uncertainty means in terms of variability
    - Standard error
    - Generating bootstrap, simulated, or hypothetical samples

Note: Response cannot simply say, “generate new samples”, as those could be real or hypothetical samples

Note: If there is a true assertion but also a misconception that is relevant to the item, then the entire response is scored as a 0. See response from raw_id 183 as an example.

Example exemplary answer:

The purpose of resampling in this simulation is to mimic taking repeated samples from the population of all Delta airlines flights in 2019, which is then used to estimate the uncertainty in the average flight delay.

item e02_calculate

Your friends agreed to use 1.85 for the Standard Deviation of trials but disagreed on the Sample Estimate.
Friend 1: *"The sample average of 6.35 minutes should be used."*
Friend 2: *"The average of the 500 bootstrap trials, 6.10 minutes, should be used."*

Which friend do you agree with? Why?

To score 1, must have the following:

- Agree with Friend 1

AND

- Something about at least one of the following:
    - Needing to use the observed data or original sample
    - Not using the results from the simulation for this purpose

---
Example exemplary answer:
I agree with Friend 1, because the average from the observed data should be used. The purpose of the bootstrap simulation is to estimate the uncertainty, not to estimate the average itself.

item e03_interpret

To answer the research question, Friend 1 wanted to report the interval they calculated. However, Friend 2 said, *"You don't need an interval to answer the research question. You can report the average of the 500 bootstrap trials."*

Do you agree with Friend 2? Why or why not?

To score 1, must have the following:

- Disagree with Friend 2

AND

- Something about at least one of the following:
    - Needing to account for uncertainty
    - What the interval provides or does for you
    - How using the average of the bootstrap is not the purpose the simulation

Note: If the answer talks about needing to evaluate the plausibility, compatibility, or validity of the average, mean, or "data", generally, then mark it as a 0, as the respondent is not actually discussing the interval in terms of the research question anymore.

Example exemplary answer:
I disagree with Friend 2. An estimate of uncertainty is needed to answer the research question and reporting the average of the bootstrap trials is not intended use of the bootstrapping simulation.

item e04_center

Suppose a third friend saw the plot of 500 trials and raised a concern.
Friend 3: *"The average of 6.10 from the bootstrap trials is different than the sample average of 6.35. When these two averages are different, this suggests the bootstrap average is more trustworthy than the original sample average."*

Do you agree with Friend 3? Why or why not?

To score 1, must have the following:

- Disagree with Friend 3

AND

- Something about at least one of the following:
    - The bootstrap averages being subject to randomness
    - The bootstrap averages being simulated averages
    - The original sample average being separate from the bootstrap simulations
    - The bootstrap simulation being based on the original sample average
    - How the two averages are equally trustworthy

Example exemplary answer:

I disagree with Friend 3. The point of bootstrapping is not to find nor evaluate the sample average. Since random resampling is involved in bootstrapping, there will likely be variation in the average of the bootstrap trials. The trustworthiness of the original sample average depends on factors unrelated to the bootstrap simulation.

item e05_trials

Consider the results from the smallest number of trials (100). Was there something gained by the other groups running more trials? If so, what did they gain and why? If nothing was gained, why not?

To score 1, must have the following:

- Something about at least one of the following:
    - More trials showing or providing more possibilities, hypothetical samples, hypothetical outcomes, or randomized data [Rubric comment: This is an emphasis on the process of the simulation, instead of the product.]
    - A better or more accurate "standard error", "standard deviation", or "margin of error" [Rubric comment: This is an emphasis on the product of the simulation, instead of the process.]

Example exemplary answer:

Yes, the extra trials give a better idea of the shape of the simulated distribution. This is due to seeing more of the possible results that might occur were we to actually resample from the population.

item e06_replicate

Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results.

Friend 1: *"We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample."*
Friend 2: *"We can also run our bootstrap simulation again with the data we already have, as that is as good as a replication of the study."*

Is the approach by Friend 2 a valid way to do a replication study? Why or why not?

To score 1, must have the following:

- Disagree with Friend 2

AND

- Something about at least one of the following:
    - Replication requiring new data
    - The results not changing by reusing the same data

Example exemplary answer:
The approach by Researcher 2 is not a valid replication study. A replication study requires new data to see if the same effect appears. Rerunning the simulation with the same data wouldn't change the results and is essentially just adding more trials to the original simulation.

item e07_confound

Suppose that only morning departure times (in the range of 5am – 12pm) were included in the sample.

Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?
To score 1, must have the following:

- State that a valid answer cannot be provided

AND

- Something about at least one of the following:
    - Confounding variables, simulation, or randomization not being able to change or correct the data
    - The data being flawed or problematic
    - The lack of representation for all times

Example exemplary answer:
The friends cannot still provide a valid answer to the research question, because the time range acts as a confounding variable in the original data. The simulation cannot correct for the confounding variable.

item e08_samples

Suppose there was concern that the size of the original sample was too small.

Did the fact that your friends ran many trials of the bootstrap simulation create a larger sample size for the study? Why or why not?

To score 1, must have the following:

- Disagree that running more trials creates a larger sample size

AND

- Something about at least one of the following:
    - Trials being different from sample size
    - Sample size being fixed
    - Sample size being the same for each trial
    - More trials simply adding more trials to the study
    - Bootstrapping not creating a larger sample size, generally

---

Example exemplary answer:

No. The number of trials and the sample size are not related. The sample size is the same for each trial, no matter how many trials are run.

item e09_device

Suppose that the flight tracker at MSP airport was found to be faulty. It consistently recorded departure times as being later than it should have.

Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

There are two different ways to score a 1.

To score a 1, response must:

- Indicate that a valid answer cannot be provided

AND

- Something about at least one of the following:
    - The confounding nature of the device
    - The data being problematic
    - The simulation not correcting a problem with the device

Alternatively, to score a 1, response must:

- Indicate that a valid answer can be provided, if the amount that the device was off was the same amount or known each time [Rubric comment: This explanation allows the device inconsistency to be taken into account.]

Example exemplary answer:

No. The friends cannot provide a valid answer to the research question, because the simulation cannot correct for incorrectly measured data.

Appendix D2: Scoring rubric final draft

Scoring rubric for Simulation Understanding in Statistical Inference and Estimation (SUSIE) instrument

This document contains the 18 items from the instrument and the guidance on how to score each of those items. Images and non-item text from the instrument are not included. Each item's scoring guide describes the content of the response that must be present for the response to be scored as a 1. This is version 04. This includes changes made on or after 4/3/2021, prior to and during the second round of scoring and classifying.

General scoring guidelines: A single response is scored as 1 or 0. A score of 1 is marked when a response meets all stated scoring criteria for the item. Blank responses are not considered responses and thus are not scored. Only responses that do not meet criteria but still contain text are marked as 0. Example exemplary responses were written by Jonathan Brown.

Section: Null-hypothesis Significance Testing

item n01_purpose

What is the purpose of randomization in the simulation (in terms of how it helps you answer the research question)?

To score 1, must have the following:

- Something about at least one of the following:
    - Assuming, producing, seeing, or modeling no difference or no effect
    - Reassigning values to subjects, regardless of original group, and comparing that to the original data [Criterion added based on ID = 02 and ID = 124.]
    - Assuming the sample all came from the same population [Criterion added based on ID = 106]

Note: If response correctly addresses at least one criterion above but also states something that is both false and relevant to the item, then that response should be scored as 0. [This is based on ID = 145, whose response contains both correct and incorrect explanations.]

Note: Referring to the real data or observed result is not needed, as the topic of the question is the purpose of the simulation from a hypothetical vs. real-world interpretation. The role of the observed data is a separate entity in this instance.

Example exemplary response:

The purpose of randomization is to see how common the observed difference in the average blood pressure reduction between the two groups would be, if there were no true difference between these groups.

item n02_center

After seeing the distribution of 500 trials, Researcher 1 offered this observation: *"Since the results center around 0, that suggests there is no difference in effect between*

fish oil and soybean oil in reality."

Do you agree with Researcher 1? Why or why not?

There are two different ways to score a 1.

To score 1, must have the following:

- Disagree with Research 1, or provide a response that indicates disagreement with Researcher 1 without necessarily explicitly stating disagreement with Researcher 1

AND

- Something about at least one of the following:
    - The results in the picture being simulated, hypothetical, or fake
    - The simulation being designed to have results centered around 0
    - The distribution resulting from values being randomly assigned to groups [Criterion added based on ID = 21]
    - The distribution resulting from values being combined into one group from which values were sampled [Criterion added based on ID = 193]
    - Needing to take the observed data into account
    - The graph not providing enough information to draw that conclusion
    - Needing to compute a p-value
    - Needing to compute a confidence interval [Criterion added based on ID = 4]

Alternatively, to a score a 1 response must:

- Meet one of the criteria above under the "Something about at least one of the following" list, without needing to explicitly acknowledge disagreement with Researcher 1. As long as the reasoning clearly shows disagreement with Researcher 1, a score of 1 can still be achieved. [Criterion added to justify 0 score for ID = 79, as they do meet one of the criteria, but it is not explicitly clear that they disagree with Researcher 1.]

Comment: The simple reasoning of "need a p-value or confidence interval" may be too imprecise, as participants may still harbor a misconception about centering around 0. However, for this first field test, such reasoning will be allowed, based on the wording of the item. Thus, scores on this item may add measurement error, necessitating item adjustments for the next iteration of the instrument.

Example exemplary response:

I disagree with Researcher 1. This is only a distribution of simulated results. It should center around 0, because that is how the randomization was designed. The observed difference needs to be taken into account to say something about the difference in reality.

item n03_calculate

The researchers agreed to calculate a p-value but disagreed on the procedure.

Researcher 1: *"Start at the average of the 500 trials, and then calculate the proportion of all trials larger than that."*

Researcher 2: *"Start at the sample average difference of 7.7, and then calculate the proportion of all trials larger than that."*

Which researcher do you agree with? Why?

There are two different ways to score a 1.

To score 1, must have the following:

- Agree with Researcher 2, or provide a response that indicates agreement with Researcher 2 without necessarily explicitly stating agreement with Researcher 2

AND

- Something about at least one of the following:
    - Needing the sample average, observed data, or observed result to calculate a p-value
    - The purpose of the research question or study is to use the sample average
    - Researcher 1 is only looking at the null hypothesis
    - The correct procedure for generating a p-value
    - That what Researcher 2 is doing is the correct procedure for a p-value
    - Falsely referring to something like "proportion" but doing so in a way that indicates choosing the observed result instead of the average of the 500 trails [Criterion added based on ID = 171]

Note: If response agrees with Researcher 1 but gives an apparently correct explanation of p-value, mark as 0, because they could be mistakenly thinking the "observed results" from the simulation are real in the case of Researcher 1.

Example exemplary response:

I agree with Researcher 2, because the p-value is the probability of seeing the sample average difference or more extreme, given the null hypothesis of no-difference between the groups.

item n04_interpret

Researcher 2 calculated a p-value of 0.014. However, Researcher 2 was unsure of how to answer the research question.

Possible answer 1: *"The low p-value indicates that an average difference of 0 is not supported. There <u>is</u> evidence for a difference between the groups."*

Possible answer 2: *"The low p-value indicates that the sample average difference of 7.7 from the data is not supported. There is <u>no</u> evidence for a difference between the groups."*

Which answer do you agree with? Why?

There are two different ways to score a 1.

To score 1, must have the following:

- Agree with Possible answer 1, or provide a response that indicates agreement with Possible answer 1 without necessarily explicitly stating agreement with Possible answer 1

AND

- Something about at least one of the following:
    - The model or assumption of no difference not being supported
    - The observed difference being unlikely to have occurred by chance
    - The null hypothesis being rejected, as long as the null hypothesis is explained or indicated to be the no-difference model or assumption
    - The results being statistically significant [Criterion added based on ID = 16]
    - Low p-values indicating a difference between groups [Criterion added based on ID = 49]

Example exemplary response:

I agree with Possible answer 1, because the low p-value means the observed group difference is unlikely to have occurred, were there no group difference in the population. This suggests evidence for a group difference.

item n05_trials

Consider the results from the smallest number of trials (100). Was there something gained by the other teams running more trials? If so, what did they gain and why? If nothing was gained, why not?

There are two ways to score a 1.

To score 1, must have the following:

- Something about at least one of the following:
    - More trials showing or providing more possibilities, hypothetical samples, hypothetical outcomes, or randomized data [Rubric comment: This is an emphasis on the process of the simulation, instead of the product.]
    - A better or more accurate or precise p-value [Rubric comment: This is an emphasis on the product of the simulation, instead of the process.]
    - A more accurate or precise shape [Criterion added based on ID = 2]
    - The hypothetical or simulated or null-hypothesis mean is more accurate [see ID = 58]
    - A more accurate or precise standard deviation [Criterion added based on ID = 176]

Note: Saying that the mean gets smaller or more accurate is a statement about something (the mean) that could be interpreted as being the real mean in the population or the data. Thus, comment on the changing mean requires some clarity about the understanding of the nature of the mean. Also, commenting on the mean, p-value, or variation getting smaller is not technically accurate nor the point of running more trials. More accuracy does not necessarily equate to measures getting smaller.

Alternatively, to score a 1:

- Response indicates that there was no or little gain

AND

- Something about at least one of the following:
    - The shape staying the same

- The p-values fluctuating or not changing
- The conclusion staying the same

Note: This alternative way to score a 1 is based on ID = 39. Even if they say nothing or little was gained, showing that they attend to the shape, p-value, or conclusion indicates that they are at least using the plots in a functional way.

Example exemplary response:

Yes, the extra trials give a better idea of the shape of the simulated distribution. This is due to seeing more of the possible results under the no-difference model.

item n06_replicate

Since this was only one study the researchers wanted to do a replication study with the same sample size, to verify their results.

Researcher 1: *"We should randomly sample another 14 participants and put them through the same study that we just did."*
Researcher 2: *"We can also run our simulation again with the data we already have, as that is as good as a replication of the study."*

Is the approach by Researcher 2 a valid way to do a replication study? Why or why not?

To score 1, must have the following:

- Disagree with Researcher 2, or provide a response that indicates disagreement with Researcher 2 without necessarily explicitly stating disagreement with Researcher 2

AND

- Something about at least one of the following:
  - Replication requiring new data or people
  - The results not changing by reusing the same data
  - Replication requiring a new study [Criterion added based on ID = 47]

Example exemplary answer:

The approach by Researcher 2 is not a valid replication study. A replication study requires new data to see if the same effect appears. Rerunning the simulation with the same data wouldn't change the results and is essentially just adding more trials to the original simulation.

item n07_confound

Suppose that all of those in the fish oil group were taking a blood pressure medication, while all of those in the soybean oil group were not.

Since each trial of the simulation randomly assigned the blood pressure values to the two groups, could the researchers still provide a valid answer to the research question? Why or why not?

There are two ways to score a 1.
To score 1, must have the following:

- State that a valid answer cannot be provided, or clearly indicate through their reasoning that a valid answer cannot be provided or that they disagree with the premise of validity

AND

- Something about at least one of the following:
    - Confounding variables
    - Simulation or randomization not being able to change or correct the data
    - The data being flawed or problematic

Alternatively, to score a 1:

- Indicate that if the blood pressure difference or medication factor is taken into account that the research question can still be validly answered. [Criterion added based on ID = 38]

Example exemplary answer:

The researchers cannot still provide a valid answer to the research question, because the medication acts as a confounding variable in the original data. The simulation cannot correct for the confounding variable, as the simulation only acts on the data as they are following the experiment.

item n08_samples

Suppose there was concern that the size of the original sample was too small.

Did the fact that the researchers ran many trials of the simulation create a larger sample size for the study? Why or why not?

To score 1, must have the following:

- Disagree that running more trials creates a larger sample size, or clearly indicate through their reasoning that the sample size is not affected by the number of trials

AND

- Something about at least one of the following:
    - Trials being different from sample size
    - Sample size being fixed
    - Sample size being the same for each trial
    - More trials simply adding more trials, increasing accuracy, or adding other things that do not implicate the sample size changing [Criterion added based on ID = 8]
    - The data coming from or being based on the original sample [Criterion added based on ID = 60]
    - A larger sample size needing more samples, data, or participants [Criterion added based on ID = 9]
    - Sample size referring to original sample and not randomization sample [Criterion added based on ID = 24]

Example exemplary answer:

No. The number of trials and the sample size are not related. The sample size is the same for each trial, no matter how many trials are run.

item n09_device

Suppose that the medical device for the soybean oil group was found to be faulty. It gave consistently higher readings than it should have.

Given the simulation that was run, could the researchers still provide a valid answer to the research question? Why or why not?

There are two different ways to score a 1.

To score a 1, response must:

- Indicate that a valid answer cannot be provided, or clearly indicate through their reasoning that a valid answer cannot be provided

AND

- Something about at least one of the following:
    - The confounding nature of the device
    - The nature of their being multiple factors that went into the measurement [Criterion added based on ID = 28]
    - The data being problematic
    - The data being skewed [Criterion added based on ID = 176]
    - The simulation not correcting a problem with the device
    - The results being problematic or invalid [Criterion added based on ID = 2]
    - The information being inaccurate [Criterion added based on ID = 190]. [Rubric comment: While this is a general answer, such a response sufficiently indicates that the participant does not endorse validity of the results due to some problem with the study, the data, or what was produced.]
    - The chance for a Type I error being increased [Criterion added based on ID = 68]
    - The results or data being biased [Criterion added based on ID = 191]

Alternatively, to score a 1, response must:

- Indicate that a valid answer can be provided, if the amount that the device was off was the same amount or known each time [Rubric comment: This explanation allows the device inconsistency to be taken into account.]

OR

- Indicate some degree of disclosure about the confounding nature of the device or study [Criterion added based on ID = 49.] [Rubric comment: This explanation allows for the participant to indicate that they recognize that the confounding nature of the device is a problem and that simulation is not correcting for it.]

Example exemplary answer:

No. The researchers cannot provide a valid answer to the research question, because the simulation cannot correct for incorrectly measured data.

Section: Estimation

item e01_purpose

What is the purpose of resampling in the bootstrap simulation (in terms of how it helps you answer the research question)?

To score 1, must have the following:

- Something about at least one of the following:
    - Mimicking or representing a process or aspect involving a population
    - Allowing the data to represent the population (without referring to increasing representativeness or external validity) [Criterion added based on ID = 16]
    - Model or show sampling variability [Criterion added based on ID = 3]
    - Estimating or quantifying variability [Criterion added based on ID = 52]
    - What estimating uncertainty means in terms of variability
    - Standard error
    - The sampling distribution [Criterion added based on ID = 15]
    - Generating bootstrap, simulated, or hypothetical samples or data
    - To simulate random sampling [Criterion added based on ID = 21]
    - To get results similar to collecting data from the population [Criterion added based on ID = 55] [Rubric comment: This answer was deemed a 1, due to the presence of the word "similar", which suggests a mimicry or representation of a resampling from the population, instead of actually resampling from the population.]

Note: Response cannot simply say, "generate new samples", as those could be real or hypothetical samples

Note: If response refers to sampling variation, response must address it correctly. E.g., saying the simulation "allows for" or "accounts for more" sampling variation should be scored a 0.

Example exemplary answer:

The purpose of resampling in this simulation is to mimic taking repeated samples from the population of all Delta airlines flights in 2019, which is then used to estimate the uncertainty in the average flight delay.

item e02_calculate

Your friends agreed to use 1.85 for the Standard Deviation of trials but disagreed on the Sample Estimate.

Friend 1: *"The sample average of 6.35 minutes should be used."*

Friend 2: *"The average of the 500 bootstrap trials, 6.10 minutes, should be used."*

Which friend do you agree with? Why?

To score 1, must have the following:

- Agree with Friend 1, or clearly indicate through their reasoning that they agree with Friend 1

AND

- Something about at least one of the following:
    - Needing to use the observed data, sample average, or original sample
    - Not using the results from the simulation for this purpose
    - Bootstrapping as taking samples from the original sample [Criterion added based on ID = 30]

Example exemplary answer:
I agree with Friend 1, because the average from the observed data should be used. The purpose of the bootstrap simulation is to estimate the uncertainty, not to estimate the average itself.

item e03_interpret

To answer the research question, Friend 1 wanted to report the interval they calculated. However, Friend 2 said, *"You don't need an interval to answer the research question. You can report the average of the 500 bootstrap trials."*

Do you agree with Friend 2? Why or why not?

To score 1, must have the following:

- Disagree with Friend 2, or indicate through their reasoning that they disagree with Friend 2

AND

- Something about at least one of the following:
    - Needing to account for uncertainty
    - What the interval provides or does for you
        - Note: The above criteria may include incorrect explanations about what the interval provides. For the purpose of this item, those incorrect explanations still allow for a 1 score. The exception is for explanations that refer to evaluative language indicative of null-hypothesis significance testing, p-values, or bias. In these evaluative cases, a score of 0 is provided, because the idea of what the purpose of the research question was clearly not interpreted correctly by the participant.
    - How using the average of the bootstrap is not the purpose the simulation
    - How only reporting the average of the trials is problematic [Criterion added based on ID = 73]
    - Needing to report an interval, generally [Criterion added based on ID = 3]
    - How use of an interval accounts for sampling error (or “errors” generally) [Criterion added based on ID = 13]
    - Something about the population mean or parameter [Criterion added based on ID = 82]

Note: If the answer talks about needing to evaluate the plausibility, compatibility, or validity of the average, mean, or “data”, generally, then mark it as a 0, as the respondent is not actually discussing the interval in terms of the research question anymore. [See ID = 31 as an example.]

Note: Due to the nature of the item, a sufficient answer is to simply say "an interval is needed". However, if more explanation is provided that is indicative of evaluative language, per the first note above, then a 0 is scored.
Note: If they correctly say that an interval is needed, but they explicitly say that they don't know why, then score as a 0.

Example exemplary answer:
I disagree with Friend 2. An estimate of uncertainty is needed to answer the research question and reporting the average of the bootstrap trials is not intended use of the bootstrapping simulation.

item e04_center

Suppose a third friend saw the plot of 500 trials and raised a concern.
Friend 3: *"The average of 6.10 from the bootstrap trials is different than the sample average of 6.35. When these two averages are different, this suggests the bootstrap average is more trustworthy than the original sample average."*

Do you agree with Friend 3? Why or why not?

To score 1, must have the following:

- Disagree with Friend 3, or indicate through their reasoning that they disagree with Friend 3

AND

- Something about at least one of the following:
    - The bootstrap averages being subject to randomness
    - The bootstrap average being based on the number of trials [Criterion added based on ID = 119]
    - The bootstrap averages being simulated averages
    - The original sample average being separate from the bootstrap simulations
    - The bootstrap simulation being based on the original sample average
    - How the two averages are equally trustworthy
    - How the mean of the trials or simulation probably won't be the same as the sample mean [Criterion added based on ID = 11]
    - Indicating that the original sample mean should be prioritized [Criterion added based on ID = 61]
    - The bootstrap simulation is intended to model variation [Criterion added based on ID = 121]
    - How trustworthiness is irrelevant for the difference in averages [Criterion added based on ID = 188]
    - How the original distribution may not necessarily be Normally distributed as the reason for the bootstrap average being different from the sample average [Criterion added based on ID = 173]

Note: It is OK if the explanation argues how the bootstrap average is less trustworthy than the sample average, though this indicates limited conceptual development about what the simulation is doing.

Example exemplary answer:

I disagree with Friend 3. The point of bootstrapping is not to find nor evaluate the sample average. Since random resampling is involved in bootstrapping, there will likely be variation in the average of the bootstrap trials. The trustworthiness of the original sample average depends on factors unrelated to the bootstrap simulation.

item e05_trials

Consider the results from the smallest number of trials (100). Was there something gained by the other groups running more trials? If so, what did they gain and why? If nothing was gained, why not?

There are two ways to score a 1.

To score 1, must have the following:

- Something about at least one of the following:
    - More trials showing or providing more possibilities, hypothetical samples, hypothetical outcomes, or randomized data [Rubric comment: This is an emphasis on the process of the simulation, instead of the product.]
    - A better or more accurate "standard error", "standard deviation", or "margin of error" [Rubric comment: This is an emphasis on the product of the simulation, instead of the process.]
    - The hypothetical, simulated, or bootstrapped mean is more accurate or more precise [Criterion added based on change in rubric for n05 and ID = 15]
    - A more accurate or precise shape [Criterion added based on change in rubric for n05]. Also, a more "reliable" shape is fine. [Criterion added based on ID = 77]
    - A more reliable estimate of the mean [Criterion added based on ID = 27]
    - Something about a better or clearer picture of the population, without referring to external validity. [Criterion added based on ID = 45] Response can refer to representation but not generalizability. [See ID = 52 for a general version of this] Response cannot refer to more representative data. [See ID = 48]
    - A better understanding about the distribution of the sample statistic [Criterion added based on ID = 147]
    - Less impact from the randomness of sampling. [Criterion added based on ID = 148]

Note: If response focuses on the mean or measure of variation getting smaller, this should be marked a 0. With more trials, both measures vacillate instead of only getting smaller, and obtaining a smaller mean or measure of variation is not the point of running more trials. Referring to accuracy, while not an ideal response, is technically a more correct interpretation.

Alternatively, to score a 1:

- Response indicates that there was no or little gain

AND

- Something about at least one of the following:
    - The shape staying the same
    - The standard error, standard deviation, or margin of error fluctuating or not changing
    - The conclusion staying the same

Note: The idea is that even if they say nothing or little was gained, showing that they attend to the shape, measures of variation, or conclusion, indicates that they are at least using the plots in a functional way.

Example exemplary answer:

Yes, the extra trials give a better idea of the shape of the simulated distribution. This is due to seeing more of the possible results that might occur were we to actually resample from the population.

item e06_replicate

Since this was only one study your friends wanted to do a replication study with the same sample size, to verify their results.

Friend 1: *"We should randomly sample another 150 departures and analyze them the same way we analyzed the original sample."*
Friend 2: *"We can also run our bootstrap simulation again with the data we already have, as that is as good as a replication of the study."*

Is the approach by Friend 2 a valid way to do a replication study? Why or why not?

To score 1, must have the following:

- Disagree with Researcher 2, or provide a response that indicates clear disagreement with Researcher 2

AND

- Something about at least one of the following:
    - Replication requiring new data or people
    - Testing results in a new setting [Criterion added based on ID = 16]
    - The results not changing by reusing the same data
    - Replication requiring a new study [Criterion added based on rubric adjustment to n06.]

- Note: Incorrectly using evaluative language, such as hypothesis testing or p-values, is not relevant to the point of this item. Thus, as long as the criteria are satisfied above, use of evaluative language is no grounds for a 0 as it would be for other items. [See ID = 112 for an example being scored as a 1.]

Example exemplary answer:

The approach by Researcher 2 is not a valid replication study. A replication study requires new data to see if the same effect appears. Rerunning the simulation with the same data wouldn't change the results and is essentially just adding more trials to the original simulation.

item e07_confound

Suppose that only morning departure times (in the range of 5am – 12pm) were included in the sample.

Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

There are two ways to score a 1.

To score 1, must have the following:

- State that a valid answer cannot be provided, or clearly indicate through their reasoning that a valid answer cannot be provided or that they disagree with the premise of validity

AND

- Something about at least one of the following:
    - Confounding variables, simulation, or randomization not being able to change or correct the data
    - The data being flawed or problematic
    - The lack of representation for all times
    - Restating the research question to address the restricted range of times that could be generalized to [Criterion added based on ID = 4]

Alternatively, to score a 1:

- Indicate that it is valid to answer a different research question focused on the 5am – 12pm time range. [Criterion added based on ID = 50]
- Indicate that it is valid if random sampling from the entire day occurred and the sample just happened to be of morning flights by chance. [Criterion added based on ID = 193]

Note: It is insufficient to only say that the data are not randomly selected. [See ID = 80]

Example exemplary answer:

The researchers cannot still provide a valid answer to the research question, because the time range acts as a confounding variable in the original data. The simulation cannot correct for the confounding variable.

item e08_samples

Suppose there was concern that the size of the original sample was too small.

Did the fact that your friends ran many trials of the bootstrap simulation create a larger sample size for the study? Why or why not?

To score 1, must have the following:

- Disagree that running more trials creates a larger sample size, or clearly indicate through their reasoning that the sample size is not affected by the number of trials.

AND
- Something about at least one of the following:
    - Trials being different from sample size
    - Sample size being fixed
    - Sample size being the same for each trial
    - More trials simply adding more trials to the study, increasing accuracy, or adding other things that do not implicate the sample size changing [Criterion added based on n08 rubric]
    - The data coming from or being based on the original sample [Criterion added based on n08 rubric]
    - A larger sample size needing more samples, data, or participants [Criterion added based on n08 rubric]
    - Sample size referring to original sample and not randomization sample [Criterion added based on n08 rubric]
    - Bootstrapping not affecting the original data [Criterion added based on ID = 148]

Example exemplary answer:
No. The number of trials and the sample size are not related. The sample size is the same for each trial, no matter how many trials are run.

item e09_device

Suppose that the flight tracker at MSP airport was found to be faulty. It consistently recorded departure times as being later than it should have.

Given the bootstrap simulation that was run, could your friends still provide a valid answer to the research question? Why or why not?

There are two different ways to score a 1.

To score a 1, response must:
- Indicate that a valid answer cannot be provided, or clearly indicate through their reasoning that a valid answer cannot be provided

AND
- Something about at least one of the following:
    - The confounding nature of the device
    - The data being problematic or impacted, generally [Criterion added based on ID = 75]
    - The simulation not correcting a problem with the device
    - The nature of their being multiple factors that went into the measurement [Criterion added based on rubric n09]
    - The data being skewed [Criterion added based on rubric n09]
    - The results being problematic or invalid [Criterion added based on rubric n09]
    - The information being inaccurate [Criterion added based on rubric n09]. [Rubric comment: While this is a general answer, such a response

sufficiently indicates that the participant does not endorse validity of the results due to some problem with the study, the data, or what was produced.]
    - The chance for a Type I error being increased [Criterion added based on rubric n09]
    - The results or data being biased [Criterion added based on rubric n09]
    - An error in the research [Criterion added based on ID = 67]

Alternatively, to score a 1, response must:

- Indicate that a valid answer can be provided, if the amount that the device was off was the same amount, known each time, or taken into account [Rubric comment: This explanation allows the device inconsistency to be taken into account.]

Example exemplary answer:

No. The researchers cannot provide a valid answer to the research question, because the simulation cannot correct for incorrectly measured data.

Appendix E: Correspondence Materials for Interviews

Appendix E1: Instructor recruitment email template

Subject: Recruiting from your [semester and class section] class

Dear [X],

I am writing to see if you might forward a message from me to your students who have completed [EPSY 3264 or EPSY 5261] in [Fall 2019 or Spring 2020].

I am developing an instrument for my dissertation research that explores student perceptions about the hypothetical nature of statistical simulations. The purpose of developing this instrument is to better understand student misconceptions specific to simulation-based statistics courses. This instrument includes open-ended questions about two specific statistical simulations. I am working with my advisers, Robert delMas and Andrew Zieffler.

I am looking for 16 volunteers who would be willing to be interviewed as I ask them to read and respond to the questions on the instrument. During the interview, the students will be asked to verbalize everything that they are thinking as they work through the instrument. This includes reading directions, explaining their answers to questions, and describing things they find confusing. This type of interview is an essential part of the instrument development, as it will help expose flaws in the instrument that should be rewritten. It will also inform scoring considerations for future use of the instrument.

Students that complete the interview will be entered into a drawing for a $30 gift card. The interview will consist of the following:
- The student and I will agree on a time to meet on Zoom.
- The student will review an information sheet explaining the study, which will also be emailed to them.
- I will read a script describing the interview procedure.
- The student will complete the instrument through Qualtrics on a computer while describing their thinking.
- The student will be thanked and entered into the drawing.

Your students' feedback will be invaluable as I work toward creating a final version of this instrument. Would you be willing to copy and paste the message below and email it to your class? If so, please add me as BCC.

Thank you for helping me with my research!
Sincerely,

Appendix E2: Participant recruitment email template

Subject: Opportunity to help in statistics education research and chance to win a $30 gift card

Hello,

I'm a PhD student at the University of Minnesota, and I'm developing a test to evaluate how students understand statistical simulations for my dissertation research. I need feedback from students to help me improve the questions, so I've asked your instructor to share this message with you.

I am looking for 16 volunteers to meet with me at a mutually convenient time on Zoom to complete what's called a "think-aloud interview." You would just work through the test while talking out loud about your thinking. You would also describe things that are unclear or confusing.

No preparation or studying for this is needed. You will be provided with an information sheet for your records explaining the study. The interview is anticipated to take 45-60 minutes. Students that complete the interview will be entered to win a $30 gift card for their participation.

If interested, please send me an email (Jonathan Brown: brow3019@umn.edu) and I will give you more information. I may not be able to include everyone that volunteers, so please contact me soon if you would like to participate!

Thanks!

Appendix E3: Information sheet

INFORMATION SHEET FOR RESEARCH

Student understanding of statistical simulations

You are invited to be in a research study exploring how students understand and interpret statistical simulations. You were selected as a possible participant, because you have enrolled in EPSY 3264: Basic and Applied Statistics or EPSY 5261: Introduction to Statistical Methods, or you have experience with statistical simulations. 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: Robert delMas, Department of Educational Psychology at the University of Minnesota.

Procedures:

If you agree to be in this study, we would ask you to do the following things:

- Complete one "think-aloud" interview over a password-protected Zoom meeting with the student investigator, Jonathan Brown.
- This interview would involve working through one version of the test while talking out loud about your thinking. You would also describe things that are unclear or confusing. No preparation or studying is needed.
- At minimum, being audio and video recorded is required for participation in this research.

Confidentiality:

In any sort of report we might publish, we will not include any information that will make it possible to identify a participant.

All data and audio/video recordings, including your typed test responses and a complete transcript of your audio recording will be stored on a password-protected computer and Shared Google Drive that only the research team has access to.

For future use, your typed test responses and a complete transcript of your audio recording without personal identifiers will be securely stored on a password-protected computer and Shared Google Drive that only the research team has access to. Future use could include future research studies or distribution of your typed responses and complete transcript of your audio recording to another investigator for future research studies without your additional informed consent.

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. 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 researchers conducting this study are: Jonathan Brown (student investigator), Robert delMas (principle investigator and student adviser), and Andrew Zieffler (co-investigator and student adviser). You may ask any questions you have now. If you have questions later, **you are encouraged** to contact Jonathan Brown (email: brow3019@umn.edu, phone: 608-443-9883) or Robert delMas (email: delma001@umn.edu, phone: 612-625-2076).

This research has been reviewed and approved by an 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 (Toll Free: 1-888-224-8636) 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 E4: Interview protocol template

Think-aloud Interview Protocol

Test access:

Thank you for meeting with me. I will first introduce the study and make sure all of your questions are answered before we formally proceed to the test. In the case where my internet fails and I freeze or disappear, you may wait while I try to rejoin or you may exit Zoom, and we can reschedule.

First, let's get you access to the test:

1. Please click the link and enter the password that appears in the Chat window.
    a. [Link for instrument version]
    b. [Password]
2. You should see a Welcome page.
3. Don't click the "next arrow" until I tell you to later on.

Explaining the study:

Next, to introduce what we'll be doing:

1. I am developing a new statistics test with the help of students such as yourself.
2. Like a regular test, you'll be presented with questions and you'll answer them by typing.

3. However, this is not about whether your answers are right or wrong. Instead, my goal is to get a better idea of how the questions are working. So, I'm asking you to *think aloud* as you answer the questions.
    a. That means to tell me EVERYTHING you are thinking as you read and answer each question.
    b. Please read ALL text and please say ALL of your thoughts out loud.
    c. 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.
    d. I'm not here to correct your thinking or guide you; so, if you ask me any questions, I will probably turn them back to you.
4. I will remind you to think aloud throughout the test. My goal is to keep you talking.
5. 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.

6. Finally, we'll do this for at most one hour, unless you finish the test before then.

Final review of Information Sheet:

1. We'll practice answering out loud in a moment. Before that, please click on the Information Sheet link on the Welcome page.
    a. This was emailed to you ahead of time, explaining the data I'll be taking, confidentiality, and so on.
    b. Please read or review this Information Sheet one more time.
    c. Let me know when you're done, and if you have any questions about participating.

2. OK, before starting, let's first practice thinking out loud.

Think-aloud Practice:

1. I will read a question, and I'd like you to think out loud as you answer it:
    a. The question is: "How many windows are there in the place where you live?**"**

[Probe as necessary]: "How did you come up with that answer?"
[Probe as necessary]: "Tell me more about that. Why did you say [ANSWER]?"

2. OK, now let's turn to the questions that we're testing.

Think-aloud Interview:

1. Before you proceed, let me turn on the Auto-Transcription and the Video Recording.

[Turn on both features in Zoom.]

    a. If the subtitles are distracting, you can click the "Hide Subtitle" option under the "Live Transcript" button.
2. OK, we are now video recording.
3. Please, proceed when ready and begin reading and thinking out loud!

- Probes will be used if the student forgets to think aloud. Probes will not be used to elicit an answer from the student. Example probes include**:**
    - "What are you thinking?"
    - "Keep talking"
    - If asked what something means, ask "What do you think it means?"

Wrap-up:

1. Thank you for your effort on this. I will stop the recording.

[Stop video recording.]

2. Once I complete interviews with all participants, I will randomly draw one of the participants for the gift card. This should happen in a few weeks. If you win, I will notify you by email.
3. If you want a report about the study results or have any other questions, reach out to me at any time!
4. If you have no further questions, then we're all done and I will end the meeting.

Appendix F: Correspondence Materials for Field Test

Appendix F1: Email template for recruiting instructors

Subject: Request to recruit participants from your [class section] class

Dear [X],

I hope your fall prep is coming along well!

Would you be willing to allow me to recruit your students in [EPSY 5261 or EPSY 3264] from this Fall 2020 semester for my dissertation research? Please review the summary information below and let me know, when convenient. If so, we can confirm logistics later. (I am hoping to collect data from December 1st – 16th, so this is not urgent.)

About the study
I am testing an instrument that explores student understanding of the hypothetical nature of statistical simulations. The purpose is to better understand misconceptions specific to simulation-based statistics courses. This instrument consists of two sections (null-hypothesis significance testing and estimation), totaling 18 open-ended questions. I am working with my advisers, Robert delMas and Andrew Zieffler.

I am aiming to get as many students as possible. Student participation will consist of clicking a link from a class-wide email and completing the instrument in Qualtrics, on their own time. (If you think the instrument could be a beneficial class activity, we can discuss this.)

Where I will need your help
1. I would like to recruit students via class-wide emails that I would need you to forward along from me. **Are you OK forwarding multiple emails from me to your class in November and December?**
2. As an incentive, I would like to offer extra credit to the students in your class who participate. **Are you open to offering your students extra credit?**

If this sounds OK, I will follow up with further details/questions to confirm the logistics. Any initial questions?

Thank you for considering!
Sincerely,

Appendix F2: Email template for confirming study details with instructors

Note: The following email template is for instructors of remote class sections. Online instructors received the same email, except a request for an in-class appearance, highlighted in gray below, was not made. The asynchronous nature of the online sections did not allow for an in-class online appearance

.

Subject: Confirming my study details: [class section]

Greetings, [instructor name]!

Thank you again for helping me with my study. Now that we're further along, I'd like to confirm details with you.

Attachments for your reference
- Three draft emails (introduction, recruitment, and reminder)
- Instrument summary AND an unformatted version of the instrument
- The study information sheet

Decisions I need from you, preferably by 11/9/2020
1. If you're still OK with extra credit, how many points and in what manner would like to award it? If you don't want to give extra credit, I will remove that information for the final draft of the attached emails.
2. Are you OK with the timeline below?
3. In my emails, I state that the test may help for exam or quiz preparation. Are you OK with me using that statement?
4. Given your remote format, is there a way I could make a 3-5 minute class appearance to encourage participation and answer questions, without disrupting your class flow? If not, I understand. If so, I would like to make this appearance after the date of the introductory email.

Note: If you think the instrument would be a good activity and you want to devote class time to have your students complete it, let me know!

Draft timeline
11/23 Send introductory email
12/1 Send main recruitment email, AND test opens
12/9 Send reminder email
12/16 Test closes at 11:59p
12/19 I send you the list of participants for purposes of awarding extra credit
Note: "sending email" just means forwarding the email to your class. I will send you a final draft of each email ON the day to send each one.

Any other questions or concerns? I hope your class is coming along well!

Best wishes,
Jonathan

Appendix F3: Participant recruitment email template

TO: Students
FROM: Jonathan M Brown, University of Minnesota
Subject: Extra credit opportunity: online test for research study

Greetings!

You are invited to take a test focused on statistical simulations. The test consists of 18 short-answer questions, which should take 30 to 45 minutes to complete. You may use any resources that you want when completing the test *except other people*.

As an incentive, if you complete the test you will receive an additional [# of points of extra credit] toward your grade in this class. This test may also help you prepare for your final exam/quiz.

I would like to use your responses as part of a research study. Your responses and the study will help improve statistics classes like the one you're taking! Your responses will be anonymous. Please review and save a copy of the attached Information Sheet for your records.

You may start the test and click away from it; your progress will be saved for up to one week. Click the link below to return to where you left off. Your responses will be recorded once you complete the test.

You may quit the study at any time. Simply close the test before completing it and your responses will be deleted after one week of inactivity.

To participate in this study, please click the following link and use the password to access to the test:
[link]
[test password]

After you complete the test, you will be taken to a separate page to provide your email for the extra credit. Your email will not be linked to your test responses.

The test and your chance to participate will expire after December 16th. Please email any questions or concerns to me at brow3019@umn.edu.

Thank you!
Jonathan

Appendix F4: Reminder email template

TO: Students
FROM: Jonathan M Brown, University of Minnesota

Hello,

This is a reminder that you were invited to take a test designed to evaluate your interpretation of statistical simulations.

To participate in this study, please click on the following link or copy and paste it into a web browser. Also, please use the following password to access the test.
[link]
[test password]

After you complete the test, you will be taken to a separate page to provide your email for the extra credit. Your email will not be linked to your test responses.

The test and your chance to participate will expire after December 16th.

If you have any questions or trouble accessing the test, please email me at brow3019@umn.edu.

Thank you!

Appendix F5: In-class additional recruitment script template

Hello students.

I'm a PhD student at the University of Minnesota, and as part of my dissertation I'm working on using a new test to explore how students understand statistical simulations. I want to use your ideas for my research, so I've asked your instructor if I could introduce my study and encourage you to participate.

I am inviting you to take a test with 18 questions about simulations. Your answers to the test will help researchers and teachers better understand how you interpret simulations and how to make improvements in statistics courses like the one you're taking.

On December 1st you will receive a recruitment email from your instructor with further instructions. To participate in the study, you will simply click a link to the test in that recruitment email. The test should take about 30 to 45 minutes to complete, which you can do on your own time on any computer or tablet. There is no time limit.

As an incentive, if you complete the test you will receive an additional [# of points of extra credit] toward your grade in this class. This test may also help you prepare for your final quiz or exam!

Since the extra credit needs to be counted toward your grade for this semester, I will only be accepting participants until December 16th 11:59pm.

Thank you for your consideration and for helping me with my research!

Appendix F6: Information sheet for field test

INFORMATION SHEET FOR RESEARCH
Student understanding of statistical simulations

You are invited to be in a research study exploring how students understand and interpret statistical simulations. You were selected as a possible participant, because you have enrolled in EPSY 3264 or EPSY 5261. We ask that you read this form and ask any questions you may have before participating in the study.

This study is being conducted by Robert delMas, Andrew Zieffler, and Jonathan Brown, Department of Educational Psychology at the University of Minnesota.

Procedures:

If you agree to be in this study, we ask that you complete one 18-question test in Qualtrics.

After submission of your answers, you will be given the option to be awarded extra credit for your participation. You will be redirected to a page separate from the test where you can enter your class section and email.

Confidentiality:

In any sort of report we might publish, we will not include any information that will make it possible to identify a participant.

Following submission of your test answers, you will be given the option to be awarded extra credit for your participation. You will be redirected to a page separate from the test where you can enter your email. Your email will be collected by the research team but will not be linked to your test responses. Your test responses will remain anonymous. Once the research team notifies your instructor that you completed the test, your email will be deleted from the data.

For future use, your typed test responses without personal identifiers will be securely stored on a password-protected computer and Shared Google Drive that only the research team has access to. Future use could include future research studies or distribution of your typed responses to another investigator for future research studies without your additional informed consent.

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. If you decide to participate, you are free to not answer any question or withdraw at any time without affecting those relationships. If you agree to participate but decide to withdraw prior to completing the test, you can simply exit the test and your responses will be automatically deleted after one week. However, if you fully complete the test, your responses cannot be deleted, as all responses are anonymous.

Contacts and Questions:

The researchers conducting this study are Jonathan Brown (student investigator), Robert delMas (principle investigator and student adviser), and Andrew Zieffler (co-investigator and student adviser). If you have questions now or later, **you are encouraged** to contact Jonathan Brown (email: brow3019@umn.edu, phone: 608-443-9883) or Robert delMas (email: delma001@umn.edu, phone: 612-625-2076).

This research has been reviewed and approved by an 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 (Toll Free: 1-888-224-8636) 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.

This information sheet is for your records.

Appendix G: Comments for Second Round of Classification

Items n01/e01 (Purpose)

n01
e01

Major issue with this item is use of the word or idea of "represent", "representation", or "representativeness". Some responses framed this from an external validity perspective, whereas other responses seemed to frame this from a more informal perspective, such as getting a better idea of what the underlying population might look like. In cases where "representation" is apparently referring to getting a better picture of the population or having a collection of simulated samples better represents the population, this was not deemed a real-world conflation. See e01 ID = 97. In cases where "representation" sounded more like the representation of a sample or increasing the representativeness of the data, this was deemed a real-world conflation. See e01 ID = 48. For highly ambiguous cases, these were labeled as inconclusive. See e01 ID = 81.

Items n02/e04 (Center)
n02

For responses exclusively about the confidence interval including 0 without other language suggesting of rwc, these were not marked as rwc. In these cases, it was not clear how participants thought about the distribution or the center of the simulation. Using the word "data" in a response automatically indicates a product-facet rwc. Several responses agreed that there is no difference but did not provide sufficiently clear reasoning to indicate how they were interpreting the graph or results. These responses were not coded as showing rwc.

e04

ID = 8 provides compelling example of simultaneously using rwc ideas while still correctly describing the simulated nature of bootstrapping: "yes because the bootstrap average was taken from a larger number of samples, and larger sample size means greater accuracy. sampling with replacement mimics possible real sampling but provides us with more data". The two sentences conflict each other.

At least one response discussed accuracy and more of something though only through "more numbers". Given the lack of clear terminology or description, this was not coded as rwc.

ID = 48 balanced simulation language with rwc language: "I agree with friend 3 due to the fact that the bootstrap average takes over 100 simulations of the data into account. The bootstrap provides a large array of data rather than basing evaluations of data off of one simulation." This was coded as the product facet due to the usage of "large array of data", despite only referring to simulations. Interestingly, the sample data appears to be framed as a single simulation.

Items n03/e02 (Calculation)
n03

Several of the panacea-based responses appear to have misinterpreted the two researchers' approaches. Some participants appear to have interpreted that Researcher 1

(the incorrect choice) somehow ran more trials than Researcher 2, even though the same simulation output was used for both researchers.

In one case (ID = 185), a clear misconception that discredits the observed result is shown, but the participant does not explain why. As a result, this was not coded as a real-world conflation, as they could be discrediting the 7.7 for many reasons, including not understanding that 7.7 is the observed result in the first place.

e02

The two-part nature of the desired response created difficulty in interpreting some responses. For example, one participant correctly explained that the observed instead of simulated data needed to be used, but they incorrectly chose Friend 2. This misplacing of the friend choice with a correct explanation was judged to be inconclusive, from a rwc coding standpoint.

The use of the word "data" was considered a reason to code a response as showing rwc, unless it is used in the phrase "bootstrap data".

One participant (ID = 162) specifically and correctly stated that the bootstrap mimicked sampling from a larger population. However, that reasoning was used as justification for choosing the bootstrap average instead of the sample average, suggesting the participant is assigning some special power or property to that mimicry. Thus, this was coded as rwc-panacea, despite the correct reasoning.

Items n04/e03 (Interpret)

n04

Many incorrect responses to this item were ambiguous, primarily due to an apparent confusion about the null hypothesis or hypothesized model were. Thus, many Participants correctly chose Answer 1 but incorrectly explained the hypothesis being judged or the conclusion to be reached. For example, Participant 178: "I agree with research 1 because the p-value indicates that there is evidence that the difference is 0." Other incorrectly chose Answer 2 but also indicated hypothesis confusion, as with Participant 182: "I agree with answer 2 because since the p-value is lower than 0.05, we can conclude that the null hypothesis is not supported and therefore there is no difference between the groups."

Participant 2 correctly chose Answer 1, but their explanation implicated a RWC by talking about the "data" and where it was on the plot.

Participant 84 offered this explanation for no difference: "Low p value means that the data is not supported." This clearly discredits the data, but it is unclear why this is the case. That is, the participant's perspective on the simulation is unclear. Thus, this was not coded as rwc, even if this was the source of the statement. This implicitly might be a Panacea facet, as the simulation may be viewed as having the power to judge the data, instead of the hypotheses. Similar reasoning may explain Participant 117: "Answer 2 because we used the sample average value to calculate the p-value so this is the value we should use to compare it with." The sample average is being judged instead of the hypotheses, though it is clear in this case that it is to ensure congruence with the value used to compute the p-value. Similarly, Participant 152: "I agree with the second answer because based on the low p value that data cannot support the research question and therefore can be rejected." The data are rejected, but it is unclear what this actually means and why.

e03

Many responses focused on using choosing the interval in order to evaluate something about the sample mean. For example, Participant 102: "No, you should use the interval to compare and then see if the sample estimate falls into that range." Participant 125 explained advantages of bootstrapping that could be referring to RWC, were they further interviewed: "Yes, bootstrapping helps us save time and avoid the cost of repeating the experiment to get other groups of sampled data."

Some responses that simply refer to the bootstrap estimating the average delay time are unclear, regarding the degree of RWC-thinking by the participant. E.g., Participant 178: "I agree with Friend 2 because the bootstrap trial estimates the average delay time."

Items n05/e05 (Trials)

n05/e05

n05

One example clearly combining all three facets was Participant 15: "…more trials equates to more accurate data being collected…." The incorrect use of "accurate" refers to Panacea, "data" to Product, and "being collected" to Process. As the value of the accuracy seemed to be the primary theme of this explanation, this was coded as Panacea.

Reduction of margin of error, variance, standard error, or standard deviation was not coded as RWC for this item, unless it was paired with the notion of a larger sample size, due to the variance measures indeed getting smaller in some cases. Additionally, decreasing uncertainty was only coded as RWC-Panacea when paired with statement about changing sample size.

Statements about an increased sample size sometimes included additional benefits as a result of the increase, as shown by Participant 89: "Yes, something was gained. Larger sample sizes give more reliable results with greater precision." In other cases, a false statement about more data or a larger sample size was paired with a correct albeit general statement about increased statement about increased accuracy. E.g., Participant 116: "Yes something was gained, you were able to collect more data which gave you a more accurate number." These types of responses were only coded as Product, as the Panacea aspect was either inconclusive or actually true but for a different reason than stated.

Participant 133 offered an interesting set of misconceptions, ultimately coded as both Product and Panacea: "The other teams obtained lower p-values, along with a mean closer to 0. We know as the sample size increase, the p-value decreases - as they have an inverse relationship in simulations."

e05

The idea of a "trial" appeared to be synonymous with "sample size" in many participants, leading to most of the RWC codes as shown. A direct example of this synonymousness is from Participant 184: "With a larger sample size (running more trials), they gained a smaller interval estimate which allows for less uncertainty."

Just indicating less variability was not enough to code for RWC, as that reasoning could come from many sources. Instead, RWC-Panacea was coded for responses arguing for less variability, less uncertainty, or similar responses, when paired with a statement about changing sample size.

Use of the word “sample” was deemed inconclusive, unless it was used in a phrase such as “sample size” or with the word “data”. Samples as a word could be referring to hypothetical or bootstrap samples.

The word “generate” only coded for RWC if enough other RWC language used. This created coding complications for Participant 40: “Yes, the more time you run the bootstrap the accurate your interval will be because you are generating more data points. As more and more samples are generated the variability between the standard Deviation decreases. Notice the first one had a very different SD than the other 3.” The use of “generate more data points” was coded as Product, leading to a Panacea code for a more accurate interval. However, the use of “samples are generated” in isolation falls outside of the Product coding criteria, even though the earlier part of the statement was coded as Product. To keep language consistent for coding, “samples are generated” was not coded as Product, and therefore, “variability…decreases” was not coded as Panacea.

Cases where the words “data”, “sample size”, and “trials” seemed to be used interchangeably, creating obfuscation and difficulty in coding. Consider Participant 99: “The other groups were able to draw information from their surplus of data, meaning that their information would therefore be more accurate. They were likely able to achieve much closer values of the average and standard deviations because there were more trials in the first place.” In the first sentence, “surplus of data” is used, which was coded as RWC-Product. However, “trials” is used in the second sentence, suggesting “trials” and “data” are synonymous in the participant’s mind. To keep coding consistent, Panacea was only applied when the benefit was listed as a direct conditional within the same sentence, instead of being implied across sentences when wording changed. In the given example for Participant 99, “more accurate information” was coded as Panacea, as it directly follows from “surplus of data”. However, “closer values of the average” was not coded as Panacea, as it did not directly follow from “surplus of data”, but rather, “trials” within the same sentence. Exception to this “within sentence” rule occurred when another sentence directly referred to a separate sentence, which connected the Panacea statement to the Product statement, as in Participant 20: “The other groups that ran more trials compared to the first who only ran 100 gained an average that is closer to the actual, true average. They were able to accomplish this simply by taking larger sample sizes. Variability decreases as sample size increases, so a closer average to the actual one was able to be obtained in doing so.” This level of coding complexity was rife throughout responses.

Items n06/e06 (Replication)

n06

Incorrect responses to this item were presumably likely to be favored toward the Process facet, given that the item directly asked about the replication process. Accordingly, incorrect responses were implicitly framed from a standpoint of affirming simulation as a real process.

Based on responses to both replication items (n06 and e06), it is unclear to the extent to which participants have a correct, consistent definition of what replication is.

Several examples of apparent misunderstanding about what replication entails. For instance, the response from Participant 62 was deemed “inconclusive”, due to correctly differentiating between simulated and actual resampling, in addition to focusing on sampling variability estimation, instead of replication: “Yes, the approach Researcher

two is doing is valid because it is mimicking the process of sampling repeatedly (rather than researcher one's approach of actually repeating the study) from the population and then we can estimate sampling variability." In contrast, Participant 147 similarly describes a hypothetical bootstrapping process but does not link it to estimating sampling variability.

One participant admitted to not understanding replication, and while they did see validity in redoing the simulation, they did not seem to harbor a RWC: "I am not familiar with the true definition of a "replication study", I believe research 1 would be the best way to preform the replication study since you are adding in more true values rather than simulation values. However, I do see validity in researcher 2 as well since it is true that different simulation means will result by running a second simulation with the data already present." This was not coded as RWC.

Responses that explain that Researcher's process is valid but clarify that the reasoning is due to the study or some factor other than the simulation were not coded as RWC. Consider Participant 83: "Yes, because the study uses random assignment, so the results would be different from the results of the first study." The emphasis on the study using random assignment appears to remove the role of the simulation for replication.

One type of response focused on the small sample size as the reason for not picking Researcher 2. Consider Participant 100: "Approach #2 is not a valid way to run the replication of a study because 14 is not enough of a sample size to confidently feel it we got an accurate representation of the general population. However, we can feel confident with the results that we received the first time would hold up due to our bootstrapping the original sample." As in similar responses, they do not address if the simulation would be a valid form of replication, were the sample size sufficiently large. These amount to agnostic responses, as they do not address the role of the simulation. Hence, these types of responses were deemed "inconclusive", regarding any RWC.

e06

Panacea Special Case also applies to this item. Consider Participant 8: "if the first study was done by true random sampling, then that data should be representative of all departures, and sampling with replacement is representative of doing another actual sample. The only reason what it might be beneficial to do another study was if there was concern for confounding variables at the time of the original study." Using the phrase "representative of doing another actual sample" does not suggest a RWC, but instead a Panacea effect, in terms of a hypothetical process replacing a real process for the real outcome of a replication.

Items n07/e07 (Confound)

n07

One theme of incorrect responses involved referring to random assignment or a description of random assignment as the reason why a valid conclusion could still be made. This type of response was not coded as RWC, due to obfuscation over whether the participant was referring to the random assignment in the study design or the mimicry of random assignment in the simulation. For example, consider Participant 26: "Yes, the researchers can still provide a valid answer to the research question. The random assignment allows the researchers to prevent bias."

The premise of the item focuses on the Panacea facet, due to directly asking about overcoming the effects of a confounding variable. Thus, incorrect responses coded as a RWC facet were predisposed to be coded as the Panacea facet.

"Randomization" was used as a keyword for a possible RWC, whereas "random sampling" and "random assignment" were not.

e07

Similar to n07, the nature of the item prompt encouraged panacea-based incorrect responses.

Items n08/e08 (Sample Size)

n08

Since n08/e08 focused on the Product facet in the item premise, incorrect responses were predisposed to be classified as the Product facet, unless the explanation clarified that the outcome was simulated in some manner.

Several incorrect responses correctly explained that the actual sample size was not changing, but that a larger sample size was being produced or replicated. Consider Participant 2: "The way the researchers ran the study did not change the actual sample size, however, the way they ran it showed a randomization of their results which in turn replicated a larger sample size." This was coded as Panacea by process of elimination. Since they discounted the change of the actual sample size, this eliminated Product. Then, between Panacea and Process, the nature of replicating a larger sample size given the premise of this item seems to focus more on the power of simulation than the process of simulation, though both are present. This is supported by the phrase "randomization of results", which was deemed to not use language indicating a real process.

e08

Items n09/e09 (Device)

n09

e09

Appendix H: Notes for Third Round of Classification

An informal set of response themes used in the third classification analysis are presented below. Themes focused on categorizing examples of responses using language indicating a conflation of the simulation with the real world, with exceptions, as some themes were used to track the presence of other aspects across responses. Each item was given its own set of themes, though some themes repeat across items. Themes vary by specificity, as each set of themes was tailored to each set of responses to a given item. Numbers correspond to participants IDs. Additional details about these notes are highlighted below:

- Some participants may have stated a theme more than once
- **Bolded** responses indicate "clear" conflations
- For select items, responses correctly disagreed with the prompt but gave a wrong explanation. Those "disagree" responses with explanations suggesting conflation are <u>underlined</u>.

n01 (Purpose)

- remove problems or improve something [that is not clearly internal or external validity; also, not things that might be associated with variability measures] (5, 6, 9, **15**, 29, 30, 33, 34, 35, **38**, **41**, **48**, **51**, 52, 53, 54, 57, 59, 60, **73**, **74**, 78, **96, 98**, **99**, 102, **105**, 112, **114**, 115, 116, 117, 118, 127, 128, 131, 133, 134, 138, 139, 143, 144, **145**, 147, 152, 153, 154, 156, 162, 163, 164, **167**, 169, 177, 178, 184)
- study design mix-up [includes references to random sampling and random assignment; discussions of values or groups] (7, **15**, 18, 23, 25, 31, 32, 33, 35, **38**, 48, 49, 50, 56, 59, 68, 76, **83**, 89, 118, 119, 120, **122**, **125**, 128, 131, 133, 147, **157,** 158, 162, **168**, 169, 171, 183)
- anything about internal validity (7, **9**, **20**, **23**, **32**, **43**, **46**, 50, 54, **56**, 68, **69**, **75**, **76**, **89**, 102**, 121**, 125, **151**, **164, 165**, 171, **174**)
- anything about external validity (**15**, 18, **20**, **21**, 49, 68, **72**, **75**, **80, 83**, **92, 94, 99**, **100**, **111**, **112**, **121, 122, 138, 157, 161, 164, 165, 189, 190**)
- anything about variability [includes "standard error", "SD", or "uncertainty"] (17, 39, **149, 167**, 191)
- replication (26, **141**)
- more stuff (39, **72**, **87**, **109**, 149, 167)

e01 (Purpose)

- remove problems or improve something [that's not clearly internal or external validity; also, not things that pertain to variability measures] (**2,** 6, 12, 26, 30, **35**, 38, 51, 69, 73, 96, **99**, **104**, **109**, 114, **116**, **126**, 131, 138, 151, **158**, 169, **185**, 187)
- test a property (26, **33, 76, 79, 109, 151, 192**)

- study design mix-up [random assignment or random sampling of data] (80, 90, 145)
- anything about internal validity (**22, 160**)
- anything about external validity (**7, 48, 58, 75, 80,** 82**, 100, 121, 129, 156, 163**)
- anything about variability [including "sampling variability", "standard error", "SD", "uncertainty" in cases that don't cast it as a problem, "error", or "confidence interval"] (12, 17, 22, 30, 39, **50**,65, 82, 96, **115**, **181**)
- replication [using the word or describing repetition of the study or experiment] (7, 44, 109)
- more stuff (2, **23**, 28, 48, **50**, 63, 72, **77**, **84**, 86, 99, 110, 114, 116, 126, 160, **167, 178**, 183, 185)

n02 (Center) [Underlined responses correctly disagreed with the prompt.]

- no difference (**7**, **11, 12, 17, 28, 30, 38, 52, 57,** 59, **62**, **67**, **73, 75, 83, 84, 89, 92, 101, 102, 132, 145**, 152, **158**, **160, 163, 164, 180, 182, 184,** 186, 190, **191**)
- other results flaws (47, 93, **110**, 122, 139, 165, 192)

e04 (Center) [Underlined responses correctly disagreed with the prompt.]

- remove problems or improve something [that is not clearly internal or external validity; also, not things that might be associated with variability measures] (**2, 12, 21, 22**, 30, **33**, **35**, **36**, 38, **39**, 44, **45**, **46, 50, 62, 63,** 72, 84, **87, 96**, **101, 109, 110, 116, 117, 120, 126**, 128, **132**, 134, **138, 151, 152, 164, 167, 168, 187, 189, 190, 193**)
- study design mix-up [random sampling and random assignment references that aren't explicitly tied to simulation] (19,)
- anything about external validity (**19**, **21, 46, 100**)
- anything about variability (that is not casting variability as an explicit problem) (129)
- replication (referring to repetition of samples; slightly different from "more stuff") (27, 62)
- more stuff (2, **8**, **33, 36**, 48, **50**, 52, 55, 56, 59, 87, 89, 99, **109**, 116, **117**, 126, 128, 129, **134,** 157, **167,** 183, 185, 186)

n03 (Calculate)

- trial accuracy (**59**, **68, 115**)
- anything about internal validity (5)
- real samples (28)
- population parameter (49)

- more stuff (179)

e02 (Calculate)

- mean confusion (2, 22, 49, 165)
- remove problems or improve something (**5, 33, 45, 46, 57, 58, 59, 61, 62**, 80, **91, 109,** 120, **134, 151,** 162, 168, **178, 180, 185, 190**)
- data confusion (19, 20, 26, 28, 35, 56, 90, 169, 185)
- anything about external validity (**80**)
- anything about variability (120, 134)
- more stuff (**33, 79**, 80, 107, **109, 116,** 120, 130, **160, 190**)

n04 (Interpret)

- data confusion (2)

e03 (Interpret) [Underlined responses correctly disagreed with the prompt.]

- remove problems or improve something (**28**, **63, 115, 162**)
- data confusion (41, 48, 115)
- more stuff (**63**)

n05 (Trials)

- remove problems or improve something [this includes references to variability measures, as they were typically framed as an improvement or benefit] (4, **5**, 9, 12, **15**, 22, **23**, **25,** 26, **31, 43**, 48, **51, 52, 55, 68, 72**¸ 89, **93,** 97, **99, 104, 127, 133**, 135, 141, 143, 144, 145**, 161,** 163, **164**, 170, **171, 174, 184**¸**185, 187**)
- anything about internal validity (**26**, **121**)
- anything about external validity (**43, 52, 60, 98, 151, 156**)
- data confusion (57)
- replication (62)
- more stuff (**5**, 9, 10, **12, 13, 21, 25**, 31, **43, 49, 51, 54, 60**, 63, **68, 89, 97, 98, 99, 102**, 104, 116, **127, 128, 133, 135, 141, 143, 144, 145, 150, 151,** 156, **160, 163, 164, 170, 182, 184,** 185)

e05 (Trials)

- remove problems or improve something [this includes references to variability measures, as they were typically framed as an improvement or benefit] (8, 12, 14, **20, 23,** 25, 26, **36**, 40, 41, **43, 47**, 48, 49, 53, 61, 62, 63, 83, 89, **93**, 99, **102**, 104,

106, 114, 115, 116, 117, **121, 129**, 135, 141, 144, **149**, 152, 165, **171, 176**, 177, **184,** 192)

- test a property (**115**)
- anything about internal validity (**26**)
- anything about external validity (**43, 75**, 115)
- mean confusion (55)
- data confusion (30, 35)
- replication (62)
- more stuff (6. **8**, 9, **12, 14, 20, 21, 23, 25, 26,** 34, **36**, 40, 41, **43, 46, 47**, 48, **49**, 60, 61, 63, **83**, 84, **85**, **89**, 99, **102**, 104, **114**, 116, **117, 126, 133, 135**, 141, 144, 145, **146, 150, 152, 153, 160, 167, 176**, 183, **184, 192,** 193)

n06 (Replication) (Underlined responses indicate correct disagreement with the prompt)

- valid but [not a conflation category] (72, 80, 84, 97, 111, 169, 192)
- keep the same (116, 151, **154**, 163, 189)
- will be different (151, 192)
- save resources (**111, 161, 178**)
- limited reason [agrees with prompt but does not explain why] (34, 67, 72, 80, 84, 90, 116, **154**, 169, **178**, 179)
- remove problems or improve something (52)
- same as new (**12, 39, 105, 111, 150, 184**)
- boot-replacement (15, 21, 22, 50, 70, 86, 99, 121, 123, 126, 136, 147, 149, **150, 184**)
- mimicry (62, 77, **99**, 147, **184**)
- study design mix-up [referring to "random assignment" or "random sampling" in isolation or as if the simulation is effectively doing either] (21, 59, 70, 123, 161, 176, 189)
- other random [but not referring to "random assignment" or "random sampling" unless referring to the original study design] (12, 39, 45, 50, 77, 94, 119, 128, 130, 151, 165, 189, 191)
- more stuff (**29,** 170, 191)
- assume representation (97)

e06 (Replication)

- valid but [not a conflation category] (8, 26, 29, 44, 54, 138, 167, 192)
- keep the same (30, 35, 46, 94, 102, 116, 117, 138, 146, **151**, 165, 172, 189)
- will be different (44, 83, 102, 117, 141, **151,** 163, 167, 178)
- save resources (**136, 151**)

- limited reason [agrees with prompt but does not explain why] (27, 34, 54, 56, 57, 101, 172, 183, 192)
- same as new (**8, 23, 63, 78, 118, 150, 168, 184**)
- boot-replacement (**8**, 10, 15, 21, 22, 27, 29, 32, 50, **58**, **62**, 70, 72, **78**, 121, 123, **127**, 128, 130, 135, **136**, 147, 149, **150**, 165, 167, **168, 184**, 189)
- mimicry (**58, 62**, 77, 98, **100, 127,** 147, 149, **168**)
- study design mix-up [referring to "random assignment" or "random sampling" in isolation or as if the simulation is effectively doing either] (21, 70, 73, 83, 121, 135)
- other random [but not referring to "random assignment" or "random sampling" unless referring to the original study design] (**8**, 12, 13, 15, **23**, 26, 33, 39, 49, **63, 136**, 141, 163, 169, 170, 176, 189, 193)
- more stuff (10, 63, 70, **99**, 126, 128, 170 178)
- anything about internal validity (**22, 184**)
- assume representation (47, **23,** 126, 167, 193)
- sufficient sample size (49, 93)

n07 (Confound)

- valid but (**109**)
- sim random assignment or experiment (**36, 64, 77, 90**)
- maybe sim random assignment or experiment (112, **132**)
- sim other random (**71, 118, 160**)
- maybe sim other random (116, 164)
- sufficient trials (**37**)

e07 (Confound)

- limited reason (38, 86, 166)
- variability or error measures (**112, 162**)
- sim other random (**160**)
- results focus (167)

n08 (Sample Size) (Underlined responses indicate correct disagreement with the prompt)

First general type of response to prompt (explicit disagreement with prompt)

- no but sim more stuff (**2**, **4**, **36,** 53, 93**, 143**, 166)
- no but remove problems or improve something (65, 108)

Second general type of response to prompt (explanations suggest correct disagreement with the prompt, even if there is explicit agreement)

- sim more stuff (33, **49,** 62, **99, 127, 149**)

Third general type of response to prompt (explicit agreement with prompt or explanation that indicates clear agreement with prompt)

- more stuff (5, **6**, **12, 16**, **22**, **25, 38, 43, 44, 48, 50, 61, 63, 70, 74**, 76, **77, 84, 87, 89, 90, 101, 111, 113, 118, 119, 128**, 130, **141, 146, 150, 152, 160, 167, 182, 183**)

  The following categories are the reasons or additional themes for those responses only coded under "more stuff":

- limited reason (5, **12**, **25, 43, 44**, 76, **84, 90, 111, 146, 152, 160, 183**)
- boot-replacement (**50, 70**)
- without replacement (**16**)
- representation (**16, 48, 70, 119**)
- pop copies sample (**74**)
- trials [not necessarily a conflation as this is a way to track which responses directly spoke to trials] (**6**, **22, 38, 48, 61, 63, 77, 87, 89, 128**, 130, **141, 167, 182**)
- sim sample many times (**150**)
- mimic population [not necessarily a conflation as this is a way to track which responses directly spoke to trials] (**77, 113**)
- average from sim (21)
- expands sim (101)
- randomize (**87, 113, 118, 183**)
- repeat experiment (**89**)
- something different (**118**)
- remove problems or improve something [not tied to the "representation" category above] (**22**, **43**, 44, 128)

e08 (Sample Size) (Underlined responses indicate correct disagreement with the prompt)

First general type of response to prompt (explicit disagreement with prompt)

- no but sim more stuff (**4**, 31, **88**, **114**, **127**)

Second general type of response to prompt (explanations suggest correct disagreement with the prompt, even if there is explicit agreement)

- sim more stuff (**55, 99, 111, 167,** 176, 184)

- ambiguous agreement general (2, 18, 132)
- ambiguous general (130)
- remove problems or improve something (**130**, 140)

Third general type of response to prompt (explicit agreement with prompt or explanation that indicates clear agreement with prompt)

- more stuff (**6**, 10, **12**, **13, 14, 16, 20, 21, 22, 25**, 30, **35, 38, 41,** 43, **44, 47, 48**, 49, **50, 61, 63, 70, 72, 74**, 84, **87, 89, 90**, 95, **109, 124, 128, 141, 147**, 151, **152**, 153, **160, 163**, 164, **169, 182, 186**)

  The following categories are the reasons or additional themes for those responses only coded under "more stuff":

- limited reason (10, **25, 35, 41, 47**, 49, **63**, 84, **128**, 151, **152**, 153, **160**, 164, **186**)
- boot-replacement (**20**, **21, 38, 44, 50, 61, 70, 72, 90, 109, 124, 141, 147,** 151)
- representation (**16, 124**)
- trials [not necessarily a conflation as this is a way to track which responses directly spoke to trials] (**6**, **12**, **13**, **22**, 43, **48**, 49, **89**, 95, **141, 182**)
- redraw many times (**14**)
- mimic population [not necessarily a conflation as this is a way to track which responses directly spoke to trials] (**20**,)
- pop copies sample (**74**)
- randomization (**87, 163**)
- random assumption or sampling (**63, 109, 169**)
- sd issue (30)
- remove problems or improve something [not tied to the "representation" category above] (**14**, **25, 47**, 152)

n09 (Device)

- valid but (**32**)
- beyond observed (**9**)
- sim random assignment (**32**)
- maybe sim random assignment or sampling (132, 160)
- maybe sim other random (39, 100)

e09 (Device)

- variability or error measures (**13**)
- rerun consistency (**28**)

- bootstrap (**120**, **128**)
- maybe sim other random (160)