STUDENTS' UNDERSTANDING AND REASONING OF SIMULATION-BASED SIGNIFICANCE TESTING

V.N. VIMAL RAO
University of Minnesota
rao00013@umn.edu

1. INTRODUCTION

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

This type of reasoning based on a chance model can be traced back over three centuries (Stigler, 2016). It is an informal version of a significance test and is an example of inferential reasoning (Zieffler et al., 2008). Formal variants of these methods, particularly Null Hypothesis Significance Testing (NHST), have come under heavy scrutiny in both practice and teaching

(Nickerson, 2000; Wasserstein et al., 2019). Some scholars have advocated that these methods should be omitted from statistical practice (e.g., Cumming, 2014), or replaced by a different approach to probability and the evaluation of hypotheses (e.g., Hoegh, 2020; Lindley, 1975). Furthermore, while such tests have traditionally formed part of the core of introductory statistics curricula, students have considerable difficulty learning this type of inferential reasoning (Castro Sotos et al., 2007). Nevertheless, inferential reasoning and statistical testing are an integral aspect of scientific thinking and reasoning (Dunbar & Fugelsang, 2005), and remain a primary learning goal for college level introductory statistics courses (GAISE, 2016).

The last 30 years have seen the emergence of simulation-based methods and software tools to teach statistical inference in an attempt to obviate students' difficulties, replacing the mathematical-based theories and probability tables of the 20th century (Rossman & Chance, 2014). These new tools have led to the development of new curricula which are, within the last decade, beginning to be rigorously evaluated. Preliminary evidence suggests that these new methods may lead to marginal improvements in students' understanding (Brown, 2019). However, despite some observed changes in the way students reason about statistical inference with simulation (Case, 2016), some evidence suggests that many students still struggle to apply, explain, and justify inferential procedures with these new methods (e.g., Noll & Kirin, 2017).

Whatever the future may hold for significance tests in the practice and teaching of statistical inference, an understanding of students' challenges and successes in learning to reason inferentially in novel simulation-based curricula may help inform future teaching, research, and practice. Thus, this paper seeks to achieve the following goals: (1) Present a short history and philosophy of classical significance testing; (2) explore the emergence of simulation-based curricula and the simulation-based approach to significance testing; and (3) review recent research on students' understanding of and reasoning about significance tests within simulation-based curricula.

2. BACKGROUND

Understanding and using the basic ideas of statistical inference is one of the key recommended learning goals for introductory level statistics courses at the post-secondary level (GAISE, 2016, Goal 7). Statistical inference refers to the act of using data from a sample to probabilistically describe a larger population (Makar & Rubin, 2009). The act of making inferences that extend beyond observed data is an inherently uncertain task, and is called the problem of induction (Henderson, 2020). While this uncertainty plagues all forms of inference, statistical inference is based on utilizing probability to reason about this uncertainty (Romeijn, 2017).

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

Significance tests are one type of statistical test of a hypothesis, in which a claim about a population is epistemically judged (Moore et al., 2013). In the classical school of statistics, this is achieved by considering the expected consequences of a claim against observed evidence. The existence of a claim about the population in the significance testing procedure often presents difficulties for students, as they must reason about the hypothetical claim, the observed evidence, and the unknown truth about a population (e.g., Vallecillos & Batanero, 1996). The next section describes the philosophy and history of evaluating a candidate statistical hypothesis against observed evidence before discussing modern simulation-based curricula developing students' understanding and reasoning about evaluating statistical hypotheses via significance tests.

2.1. PHILOSOPHY OF HYPOTHESIS EVALUATION

The underlying philosophical theory detailing the relationship of evidence and hypotheses is called Confirmation Theory (Romeijn, 2017). Confirmation entails both positive affirmation of a hypothesis as well as disconfirmation, or a confirmation of a hypothesis's negation. There are generally three approaches to confirmation: confirmation by instances, hypothetico-deductive confirmation, and Bayesian confirmation (Norton, 2005). Confirmation by instances is the oldest of the three approaches and was greatly influenced by Carl Hempel. Acknowledging the impossibility of induction to prove universal truths, Hempelian confirmation instead focuses on the development of a hypothesis relative to observed evidence (Hempel, 1945). Observed evidence is said to Hempel-confirm a hypothesis if and only if the evidence can be considered an instance of or consistent with the hypothesis. For example, observing an individual who is a human and has 10 fingers Hempel-confirms the hypothesis that 'All humans have 10 fingers', while observing an individual who is a human and does not have 10 fingers Hempel-disconfirms the hypothesis.

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

Returning to the previous example, the hypothesis ‘All humans have 10 fingers’ implies that an observed human must have 10 fingers, and thus observing such an individual hypothetico-deductively confirms the hypothesis. While the results in this case are the same as in Hempel-confirmation, the underlying logic is fundamentally different. Yet, both Hempel confirmation and hypothetico-deductive confirmation depend on formal logic (and suffer from many logical paradoxes), and neither explicitly addresses the uncertainty in inferences through attributing probabilities to either observed evidence or hypotheses.

Bayesian confirmation, on the other hand, explicitly aims to assign probability values to hypotheses (Talbott, 2016). Bayesian confirmation is based on calculating the probability of a hypothesis, i.e., $P(H)$, as well as the probability of that hypothesis conditioned on observing some evidence, i.e., $P(H|e)$. A hypothesis is Bayes-confirmed by this observed evidence if and only if $P(H|e) > P(H)$, while it is Bayes-disconfirmed by the evidence if and only if $P(H|e) < P(H)$. In other words, if in light of the observed evidence the probability of a hypothesis increases from its previous state, the evidence Bayes-confirms the hypothesis. Although Bayesian confirmation does not suffer from the same logical paradoxes that plague Hempel confirmation and hypothetico-deductive confirmation, it is dependent on choosing an interpretation of probability that allows probability to be assigned to statements and hypothesis.

There are generally four recognized philosophical interpretations of probability: relative frequency, propensity, logical, and subjective (Hájek, 2019). In the relative frequency interpretation, probability is defined as the limit of the relative frequency of a repeatable event (e.g., von Mises, 1957). This interpretation is the most restrictive interpretation of probability, as the relative frequency is undefined for non-repeatable events and is not used to assign probabilities to statements. For example, the probability that a fair coin lands head when flipped is well-defined, while the probability that a particular vase will break when dropped is not, as dropping a single vase is not a repeatable event (after the first occurrence of it breaking).

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

Both the relative frequency and propensity interpretations are considered physical probabilities, as they are physical characteristics of an event such as a ‘coin flip’ or a ‘vase drop'. In contrast, the logical and subjective interpretations are considered epistemic, dealing with what individuals or communities of people believe or know, and are used to apply probability to statements.

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

As statistical inference utilizes probability to reason about the uncertainty of inferences, the choice between the physical interpretations of probability and the epistemic interpretations of probability fundamentally leads to different types of reasoning about the relationship between observed evidence and a hypothetical claim. Classical statistics was heavily influenced by scholars who subscribed to a physical interpretation of probability (e.g., R. A. Fisher, J. Neyman, A. Wald; Romeijn, 2017). Since physical interpretations only assign probabilities to events, these scholars necessarily adopted an approach akin to hypothetico-deductive confirmation, taking a hypothesis as given and reasoning about the consequences of that hypothesis in terms of the probabilities the hypothesis ascribed to events. Furthermore, this required that observed evidence be treated as one event from an infinite collection based on a repeatable process, thus allowing it to be ascribed a probability of occurring. A hypothesis could then be made about the probability of such an event (e.g., Neyman, 1937). On this basis, an inference could be made about the hypothesis, and quantitatively described in terms of the quality of or support for the inference with notions such as significance, likelihood, or confidence. Significance testing, one of the first approaches to evaluating statistical hypotheses, concerned itself with the probabilistic consequences of a given hypothesis, and whether observed evidence deviated significantly from a hypothesis’s probabilistically expected events.

2.2. CLASSICAL SIGNIFICANCE TESTING

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

Two centuries later, such assumptions were codified as null hypotheses, their associated probability models as null models, and the probability calculations as *p*-values (Fisher, 1925, 1935). While Fisher was not the first person to write about significance testing (e.g., Pearson, 1900), the popularity of Fisher's 1925 book *Statistical Methods for Research Workers* and his 1935 book *The Design of Experiments* spread these concepts to a wide audience beyond the realm of statisticians and popularized their use (Stanley, 1966). Fisher (1935) defined the null hypothesis as the basis for the specification of a probability distribution, and that this probability distribution would in turn serve as the basis for a significance test. To Fisher, null hypotheses were a characteristic of all experiments, and experiments existed solely to give evidence a chance at disproving a null hypothesis. The experiment functioned as a repeatable event, often with a component random process, allowing a physical interpretation of probability to be assigned to possible outcomes of the experiment, which could then be interpreted in a probabilistic hypothetico-deductive disconfirmation of the null hypothesis.

For a hypothesis to qualify as a null hypothesis, Fisher stipulated that it must be exact in its specification of a probability distribution which could subsequently be used to create an exact statistical criterion. The probability distribution represents “the frequencies with which the different results of our experiment shall occur” (Fisher, 1935, p. 190). The statistical criterion, or significance level, demarcates the threshold beyond which observed evidence would, in relation to the null hypothesis, present a logical disjunction – either an extraordinary coincidence has occurred or the hypothesis is likely incorrect (Fisher, 1956). Observed evidence laying beyond the significance level thus constituted “rational grounds for the disbelief it engenders [in the null hypothesis]” (Fisher, 1956, p. 43). Any hypothesis meeting the criteria for exactness could be chosen and given a chance to be disproved by this procedure. Thus, to Fisher, the null hypothesis was simply the hypothesis to be nullified via experimentation (Cohen, 1994).

Fisher alternately referred to the probability distribution specified by the null hypothesis as the “random sampling distribution on the null hypothesis” (Fisher, 1935, p. 62) and the “sampling distribution completely determined by the null hypothesis” (Fisher, 1935, p. 192). In this way, the existence of a probability distribution is exactly the characteristic that makes a candidate hypothesis a *null* hypothesis. These null probability distributions, or null models, thus provide the means to establish statistical criteria with which to compare observed data to the null hypothesis, all in the service of the potential nullification of a hypothesis through experimentation.

2.3. ALTERNATE APPROACHES

A significance test taking a hypothesis as given and considering evidence that the hypothesis implies is only one approach to statistical testing. Recall that Hempel-confirmation takes the observed evidence as given and considers whether the evidence contributes to the development of a hypothesis. One can consider the plausibility or credibility of potential values of a parameter in light of some observed sample statistic. Hempel called this an inductive-statistical explanation, while modern scholars consider plausibility by using what is known as a likelihood function (Barnard, 1967). This approach is an example of one of the major schools of thought for statistical inference, known as the Likelihood school. The other two most popular schools are the classical school (to which Fisher subscribed) and the Bayesian school (Bandyopadhyay & Forster, 2010).

The Bayesian school fundamentally differs from both the Likelihood school and the classical school by relying on an epistemic definition of probability and assigning probabilities to hypotheses, i.e., either a logical or subjective interpretation. The probability of the hypothesis prior to the collection of new evidence is called the prior probability, and the probability conditioned on the new observed evidence is called the posterior probability. In this manner, parameter values with posterior probabilities less than their prior probabilities are Bayes-disconfirmed. The Likelihood school also differs from the classical school by emphasizing the likelihood function, and conducting both estimation and testing procedures by evaluating the entire likelihood function based on the observed evidence (Edwards, 1972).

Not only are there multiple approaches to the evaluation of a single hypothesis in light of observed evidence, but each school of statistics has a different approach to generating decision rules for selecting between competing hypotheses. Within the classical school, the hypothesis testing approach is most common and seeks to minimize errors associated with incorrectly choosing one hypothesis over another. This is done by specifying decision criteria based on comparisons of the probability of observing events given each candidate hypothesis (Neyman & Pearson, 1928). In the likelihood school, the likelihood ratio test is used to identify whether two models differ by comparing their likelihood functions (Edwards, 1972). In the Bayesian school, the Bayes factor, the ratio of the posterior probability of two hypotheses, can be used to select the hypothesis with the higher probability (Jeffreys, 1961). A classical hypothesis test can be considered a before-data decision rule, specified without incorporating observed evidence, whereas the likelihood and Bayesian approaches explicitly include the observed data in the selection of one of the competing hypotheses through the likelihood function or posterior probability distribution, which itself is a function of the likelihood function (Hacking, 1965).

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

2.4. STUDENTS’ STRUGGLES AND CRITICISM

To create a method that both specified a decision rule and made epistemic judgements, null hypothesis significance testing (NHST) fused the decision error minimization approach with the significant difference approach. Perhaps unsurprisingly, this has led to much confusion among students, statisticians, and textbook writers (Gigerenzer, 2004). The underlying logic of

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

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

(1) Misunderstanding the logic of the test, i.e., the conditional nature of the hypothetico-deductive approach, by incorrectly conditioning on the observed evidence;
(2) Difficulty specifying hypotheses and conflating hypotheses with decision rules;
(3) Misinterpreting *p*-values as the strength of an effect;
(4) Misunderstanding the inherent uncertainty in inference and viewing test results as deterministic; and
(5) Conflating statistical and practical significance.

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

Many of the recommendations Nickerson (2000) found in the literature harken back to the core logic and procedures espoused by early 20th century statisticians such as Fisher and Neyman but forgotten through the decades. For example, using non-nil null hypotheses or providing specific alternative hypotheses were requirements in significance testing and hypothesis testing respectively as originally specified, but fell out of use in NHST. Further recommendations advocate distinguishing between the substantive contextual research question and the statistical hypothesis, re-emphasizing the role that a researchers' intuition plays in inference, just as it was emphasized by Fisher (1925) and Neyman and Pearson (1928). Some critiques go further at striking at the underlying philosophy and recommend Bayesian approaches or emphasize abduction through likelihood-based inference (e.g., Rozeboom, 1997). Despite these disagreements about what to teach and how to teach it, only recently have calls been made to eliminate statistical procedures for the evaluation of hypotheses, instead focusing entirely on estimation (Cumming, 2014). Nevertheless, significance testing and hypothesis testing continue to be a central part of statistical inference and a key learning goal in modern introductory level curricula (GAISE, 2016). Furthermore, the core purpose of significance testing, to probabilistically reason about the relation between a candidate hypothesis and observed evidence, is fundamental to all statistical inference across all schools of thought and all interpretations of probability.

2.5. SIMULATION-BASED SIGNIFICANCE TESTING

With the advent of modern computing, statistics educators began calling for the use of simulation in the classroom to supplement or even replace the traditional mathematical aspects of statistical inference which students often found difficult to comprehend (e.g., Glencross, 1988). Null models in Fisher's day took the form of probability models specified mathematically (e.g., Z, T, χ^2, F). After selecting the appropriate probability distribution as a null model, the significance level (i.e., the threshold beyond which observed evidence would, in relation to the null hypothesis, present a logical disjunction) could be identified by referring to regularly published tables or via manual calculations.

By the late 20th century, automatic hand-held calculators replaced probability tables (Moore, 1992), and some statistics educators saw in simulation-based methods an opportunity to reconsider the way students are taught the core ideas of statistical inference (e.g., Ernst, 2004). In a paper denouncing the then consensus introductory statistics curriculum as *obfuscatory*, *costly*, and *fraudulent*, Cobb (2007) articulated a set of core principles known as the three R's – "Randomize data production, Repeat by simulation to see what's typical, and Reject any model that puts your data in its tail" (p. 13). These principles served as a basis for what is known as simulation-based inference in introductory statistics curricula (Rossman & Chance, 2014).

Simulation-based inference generally refers to the use of a statistical model defined by a simulator to perform statistical inference (Cranmer et al., 2020). A simulator is any tool that can enact simulation, or in a statistical context, any computational algorithm that defines a population and assumes a data generating process to randomly generate multiple sets of sample data (Carsey & Harden, 2014). Simulators for statistical inference utilize resampling, i.e., a statistical procedure that reuses data from an observed sample in the service of statistical inference (Chernick, 2012). While there are many types of resampling techniques (e.g., bootstrapping, jackknifing, cross-validation, and randomization), introductory curricula overwhelmingly focus on bootstrapping and randomization (Brown, 2019). A bootstrap resampling procedure iteratively samples with replacement from the original dataset until the original sample size is reached. This process mimics the process of random sampling from a population, with the original sample operating as an estimate of the population distribution. A randomization resampling procedure rearranges or regroups observations from the original dataset, in order to mimic the process of random assignment into experimental groups. Compared to its mathematical predecessor, simulation-based inference generates an approximation of the null distribution via simulation rather than an approximation based on a mathematically derived theoretical probability distribution.

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

Before long, collections of activities became entire curricula, and simulation applications and software were either adopted or specifically designed for a curriculum. The Rossman-Chance applets initially developed by Chance and Rossman (2006) were incorporated into the *Introduction to Statistical Investigations* curriculum (ISI; Tintle et al., 2020), the *Change Agents for Teaching and Learning Statistics* (CATALST) curriculum was developed around the TinkerPlots software (Konold & Miller, 2005; Zieffler et al., 2019), and StatKey was designed specifically for the *Statistics: UnLOCKing the Power of Data* curriculum (Lock5; Lock et al., 2021; Morgan et al., 2014).

These simulation software tools vary greatly in the degree to which users specify the characteristics of a simulator. For example, TinkerPlots requires users to construct models using a variety of resampling devises such as mixers and spinners (see Justice et al., 2018, for a detailed explanation of the user interface of TinkerPlots), whereas StatKey requires users to select one of several models identified by their functionality (e.g., Bootstrap Confidence Interval for a Single Mean, Randomization Hypothesis Test for a Difference in Proportions). Furthermore, each curriculum approaches simulation-based inference in a different way: ISI focuses on randomization before connecting simulation to mathematical-based methods (Tintle et al., 2011); Lock5 focuses on bootstrapping before introducing randomization and ultimately making the connection to mathematical-based methods (Lock et al., 2021); and CATALST focuses on model creation through model eliciting activities before introducing randomization (and notably does not introduce students to mathematical-based methods; Garfield et al., 2012; Justice et al., 2020). Despite their differences, all three curricula emphasize the role of simulators in statistical inference and utilize active learning methods to promote student learning.

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

2.6. SUMMARY AND PROBLEM STATEMENT

The classical approach to significance testing utilizing null hypotheses, null models, and *p*-values have long drawn criticism from practitioners and theorists alike (Cohen, 1994). Debates over their utility and appropriateness have occurred almost continually since their formalization in the early 20th century through the present day (e.g., Berkson, 1938; Gigerenzer, 1993; Hogben, 1957; Morrison & Henkel, 1970; Nickerson, 2000; Wasserstein et al., 2019). Beyond students' difficulty in learning the method, much of the critique of the classical approach to significance testing has been centered around either its process of drawing conclusions or its underlying logic (e.g., Gigerenzer, 2004; Rozeboom, 1997). Despite such controversies, significance testing continues to be a central part of statistical inference and a key learning goal in modern introductory level curricula, including simulation-based curricula (GAISE, 2016).

With increased training efforts and a nascent evidence basis, simulation-based methods appear to be ascendent as a pedagogical tool for introductory level statistics. The core difference between these simulation-based approaches to significance testing and their mathematical-based predecessors is a simulator representing a data generating process under a null hypothesis (Cobb, 2007; Fisher, 1935). The key feature that allows a null hypothesis to be tested is its simulator that specifies an underlying null model. As more simulation-based curricula and software tools are developed, and as calls for the reform of statistics instruction explicitly recommend the elimination of significance testing, there is a need for research examining students' understanding of significance tests in simulation-based curricula. In particular, (1) to what extent do students' difficulties, documented within mathematical-based curricula, persist even in simulation-based curricula, (2) what is students' understanding of null model simulators, and (3) what if any unique aspects to students' reasoning about significance testing emerge with simulation-based methods. This paper reviews the current body of empirical evidence related to this topic.

3. EMPIRICAL EVIDENCE

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

3.1. STUDENTS' UNDERSTANDING OF SIMULATION-BASED SIGNIFICANCE TESTING

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

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

This section next explores on how simulation-based inference affects students' understanding of significance testing.

3.1.1. Students' Understanding of Null Models

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

Table 1

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

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

*Exact number of student respondents per group was not reported, and instead are estimated based on a total of 1116 students across 34 simulation-based sections and 2 traditional sections.

^Results by topic only reported for students scoring less than 40% overall.

Hildreth et al. (2018) found a similar pattern when comparing the students' scores from sections utilizing the CATALST curriculum and the Lock5 curriculum to two traditional mathematical-based curricula – the average posttest score on three items from CAOS measuring understanding of *p*-values for 1584 students in traditional curricula was 62.8% (see Appendix A for more about these CAOS items), while the average posttest score on the same items was 82.1% for 770 students in the CATALST curriculum and 83.8% for 758 students in the Lock5 curriculum. Similar results in overall comparisons between curricula were found by Garfield et al. (2012) when comparing the CATALST curriculum to a traditional mathematical-based curriculum. Garfield et al. utilized the Goals Outcomes Associates with Learning Statistics assessment (GOALS), based on sixteen items from CAOS but with an additional seven items explicitly focusing on the use of simulation methods to draw inferences. Average student scores were higher for students in the CATALST curriculum, and were also on average higher for the seven simulation-based inference items compared to three items related to confidence intervals and four items related to sampling variability.

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

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

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

Table 2

Randomization test plan with examples (Frischemeier & Biehler, 2013)

No.	Step	Example solution to the ESP task (Rossman et al., 2001)	Expected solution to the Muffins task (Biehler et al., 2003)
1	**Observation** Which difference do you observe between the means of the two groups in the dataset?	Number of correct answers = 20	Mean of Time_Reading of boys = 2.685 Mean of Time_Reading of girls = 3.503 Difference = 0.818
2	**Hypothesis** As said in the task, the difference of the means of the two groups could have occurred at random. Generate an adequate Null Hypothesis for your investigation.	The person does not have extrasensory perception (ESP). He/She guesses with a success rate p = 0.25.	The difference of the means of Time_Reading of boys and girls has occurred at random.
3	**Simulation of H0** How can you investigate the null hypothesis with a simulation? Explain your procedure.	Drawing 40 times with replacement from an urn which is filled with 4 balls: 1 ball is labeled "right" and 3 balls are labeled "false".	Place the 533 cases of Time_Reading in urn1. Construct urn2 with 232 balls labeled "male" and 301 balls labeled "female". Draw 533 times without replacement.
4	**Test Statistic** Define the test statistic.	X = number of correct predictions	X = mean of group 1 minus mean of group 2
5	***p*-value** Calculate the *p*-value	P(X>20) = 0.0004	P(X>0.818) = 0.0006
6	**Conclusions** Which conclusions can you make regarding your null hypothesis?	The *p*-value is very small, so we have strong evidence against our null hypothesis. We assume that the fortune teller has not guessed. Another possibility is: he could have guessed but that would have been very unlikely.	The *p*-value is very small, so we have strong evidence against our null hypothesis. Another possibility is: the difference occurred at random, but that is very unlikely.

Taken together this evidence suggests that understanding how to draw conclusions from significance tests may not imply understanding the role of null models in significance tests nor the role of simulators in simulation-based tests, and the relationship of *p*-values with the underlying null models. However, such an interpretation of this evidence is tenuous at best – inferences based on students responses to different GOALS-4 items depends on the marginal reliability and distinctness of these items when measuring differences in students' understanding, and inferences based on the frequency of correct responses to written tasks depend on the quality of the rubric distinguishing between correct and incorrect responses. Nevertheless, differences in average correct responses on each item in GOALS-4 and the varying proportion of correct responses according to Frischemeier & Biehler's (2013) randomization test plan highlight that there *may* be gaps in students' understanding. Yet, neither study provides evidence of students' reasoning and thinking that may identify which of the parts of conducting a significance test they may have struggles with, or how they conceptualize the task of conducting a significance test and if it at all differs from researchers' expectations.

3.1.2. Students' Creation of Null Model Simulators

Some students' difficulties in conducting simulation-based significance tests may be explained by students' struggles in utilizing simulation software such as TinkerPlots to model the exact characteristics of a null hypothesis and its underlying null model. To further explore how students understand null models, both in terms of its function statistically as well as how to utilize the TinkerPlots software to successfully create a null model simulator, Biehler et al. (2015) examined students' submitted written work from 18 pairs of pre-service mathematics teachers recruited two months after they completed an introductory statistics course. These students were then given the Verdienststrukturerhebung [Structure of Earnings Survey] (VSE) task based on data collected by the German Statistisches Bundesamt [Statistics Bureau] (Figure 1). Students were asked to fill out a blank randomization test scheme, a slightly modified version of Frischemeier and Biehler's (2013) test plan, and their answers were rated as 'successful' if they adhered to expected solutions.

Figure 1

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

> In the dataset you can see the monthly salaries of 861 women and men of the year 2006. The display suggests that women are way behind men concerning their salary. Someone argues against the result of the group comparison between women and men that only 861 employees were asked. Therefore, the differences could have emerged due to the selection of our sample.
>
> **YOUR TASK**: Now check if there is evidence against the assumption that there is no difference between women and mean in the population with regard to their average salary. (This would mean that we can expect similar differences for all employees.)

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

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

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

However, students appeared to struggle with determining whether the device should be set to 'with replacement' or 'without replacement'. Two groups of students who correctly chose 'without replacement' simply compared this task to a previous activity, without explicitly acknowledging that this selection allowed the TinkerPlots device to mimic the random allocation process. The two groups who incorrectly chose ‘with replacement’ did so for different reasons, with one group desiring to maintain the same chance of improving for each individual in the device, and the other group hoped to model a bootstrap resampling process to facilitate generalizations of their results, despite this not representing the study design of the problem.

Figure 2

The Dolphin Therapy problem (Noll & Kirin, 2017)

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

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

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

Noll and Kirin (2017) also noted that students struggled to operationalize the null hypothesis’s statement of ‘no differences’ at the group level and instead modeled an equal chance of improving or not improving for each individual. This struggle was also noted by Biehler et al. (2015), despite students being able to correctly specify a null hypothesis utilizing mathematical symbols. However, these struggles may simply be due to the idiosyncratic features of specifying ‘replacement’ in TinkerPlots, which may be unintuitive for students as they learn to conduct significance tests and conceptualize null models. Nevertheless, together these two studies suggest students may struggle to utilize TinkerPlots to transcribe null hypotheses into null model simulators. Furthermore, the correct specification of a null hypothesis with mathematical symbols does not appear to imply that students can correctly operationalize the null hypothesis as a null model.

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

Figure 3

Facebook task from MOST (Noll & Kirin, 2016)

> Facebook is a social networking Web site. One piece of data that members of Facebook often report is their relationship status: single, in a relationship, married, it's complicated, etc. With the help of Lee Byron of Facebook, David McCandless - a London-based author, writer, and designer - examined changes in peoples' relationship status, in particular, breakups. A plot of the results showed that there were repeated peaks on Mondays, a day that seems to be of higher risk for breakups.
>
> Consider a random sample of 50 breakups reported on Facebook within the last year. Of these 50, 20% occurred on Monday. Explain how you could determine whether this result would be surprising if there really is no difference in the chance for relationship break-ups among the seven days. *(Be sure to give enough detail that someone else could easily follow your explanation.)*

Building on the findings of Noll and Kirin (2016), Noll et al. (2016) found evidence that students generally understand the role that a null model simulator plays in simulation-based inference. Analyzing the Facebook task again, along with the Music Note task (Figure 4), Noll et al. (2016) examined students' explanations in terms of how they related to four conjectured phases of inferential reasoning. Noll et al. (2016) hypothesized that students' thinking occurred in four phases in which students: (a) appropriately construct a TinkerPlots model corresponding to the null hypothesis; (b) use the model to generate a single trial and suitably represent its outcome; (c) generate multiple trials to create a distribution; and (d) utilize the results from all three previous phases to draw conclusions. Noll et al. found that even when students struggled to construct an appropriate TinkerPlots model corresponding to the null hypothesis, they were able to enact simulations and reason about the sampling distribution correctly. Two common errors were incorrectly assuming that 'by chance' implies a probability of 0.5 and designing a null model simulator based on the observed results rather than the specifications of a null hypothesis.

Figure 4

The Music Note task from MOST (Garfield et al., 2012)

> Some people who have a good ear for music can identify the notes they hear when music is played. One note identification test consists of a music teacher choosing one of the seven notes (A, B, C, D, E, F, or G) at random and playing it on a piano. The student is standing in the room facing away from the piano so that they cannot see which note the teacher plays on the piano. The note identification test has the music student identify 10 such notes.
>
> This note identification test was given to a young music student to determine whether or not the student has this ability. The student correctly identifies 7 notes out of the 10 that were played. Explain how you would use what you learned in this class to determine how surprising this result is and whether it is strong evidence that the student has the musical ability to accurately identify notes? *(Be sure to give enough detail that someone else could easily follow your explanation.)*

It should be noted that the Facebook task and Music Note task are both theoretically isomorphic and have a substantially different study design than the Dolphin Therapy Task, which requires different settings in TinkerPlots. Specifically, students do not have to specify 'without replacement' for the Facebook task, a setting which proved difficult for students studied by Noll and Kirin (2017). Therefore, students' ability to utilize TinkerPlots to realize their intended models may be limited both by gaps in their understanding of the full TinkerPlots functionality as well as gaps in understanding null hypotheses across various contexts. Similarly, their understanding of the role of simulation and their ability to enact simulation may only be limited to situations in which the resampling method required is bootstrap resampling, as is the case in the Facebook task and Music Note task. For example, one group of students studied by Noll and Kirin (2017) incorrectly attempted to utilize bootstrap resampling in the Dolphin Therapy task, despite the problem requiring randomization resampling, manifested by students' difficulties specifying 'without replacement' in TinkerPlots. Nevertheless, it appears that at least in some scenarios and tasks students are able to successfully create null simulators with TinkerPlots.

3.1.3. Summary

Preliminary evidence from evaluations of students in simulation-based curricula suggest that simulation-based inference may improve students' understanding of significance testing compared to mathematical-based inference, particularly in terms of interpreting *p*-values (e.g., Hildreth et al., 2018; VanderStoep et al., 2018). Yet, an understanding of the conclusions that can be drawn from a significance test may not imply that students also understand the role that null models play in significance testing (Frischemeier & Biehler, 2013; Sabbag et al., 2015). While students struggle to create null model simulators in TinkerPlots (Biehler et al., 2015; Noll & Kirin, 2017), for some problems they seem generally able to use TinkerPlots to represent their intended models (Noll & Kirin, 2016). Furthermore, within some problem contexts, students seem to generally understand the role of simulation with regard to the null model simulator and significance testing (Noll et al., 2016). Therefore, students' struggles with simulation-based significance testing may be due to idiosyncrasies of particular problem contexts and types of statistical study designs, or perhaps due to difficulties extracting a null hypothesis from context, converting the null hypothesis into a specific null model, and ensuring that the null model contains complete information to facilitate the creation of a null model simulator in TinkerPlots.

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

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

3.2. STUDENTS’ REASONING ABOUT NULL MODEL SIMULATORS

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

Figure 5

The NFL task (Noll et al., 2018b)

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

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

One study by Noll et al. (2018a) suggests that an inherent human predisposition for narrative sensemaking may explain how students construct null model simulators as well as how they interpret them. In general, narrative reasoning processes can help students organize information into a coherent structure (e.g., Clark & Rossiter, 2008). Furthermore, statistical models may be inherently narrative, as they "bring forth important aspects of a problem, contain an underlying statistical structure of a process, and are purposeful (used to make sense of a problem)" (Noll et al., 2018a, p. 1269).

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

Even when students are able to successfully create a null model simulator, they may not understand the random source of variation it represents. A study of in-service statistics teachers by Justice et al. (2018) utilized structured interviews to examine teachers' understanding of null model simulators as a data generating process (DGP). A DGP approach to null model simulators exemplifies Konold et al.'s (2007) theory of understanding distributions through modeling and the core principle of randomization as a data production process (Cobb, 2007; Fisher, 1935). The teachers in the study explicitly focused on creating null model simulators that replicated the manner in which the original data was produced, emphasizing some elements such as the temporal sequence of the study that do not affect the simulation results. In doing so, Justice et al.'s participants unanimously viewed the null model simulator's role as facilitating comparisons between the hypothesis and the evidence that could lead to an inference or conclusion. While this comparison is fundamental to the task of a confirmation theoretic significance test, it emphasizes the null model simulator's by-product (i.e., the simulated sampling distribution) rather than the random data generating process that it represents when conceptualizing the null model simulator's core purpose and role in significance testing. But, while these teachers teach the CATALST curriculum, they were not trained with it as students, and their understanding of null model simulators in relation to the sampling distribution may be a residual effect of their mathematical-based training.

3.3. SUMMARY

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

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

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

4. DISCUSSION

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

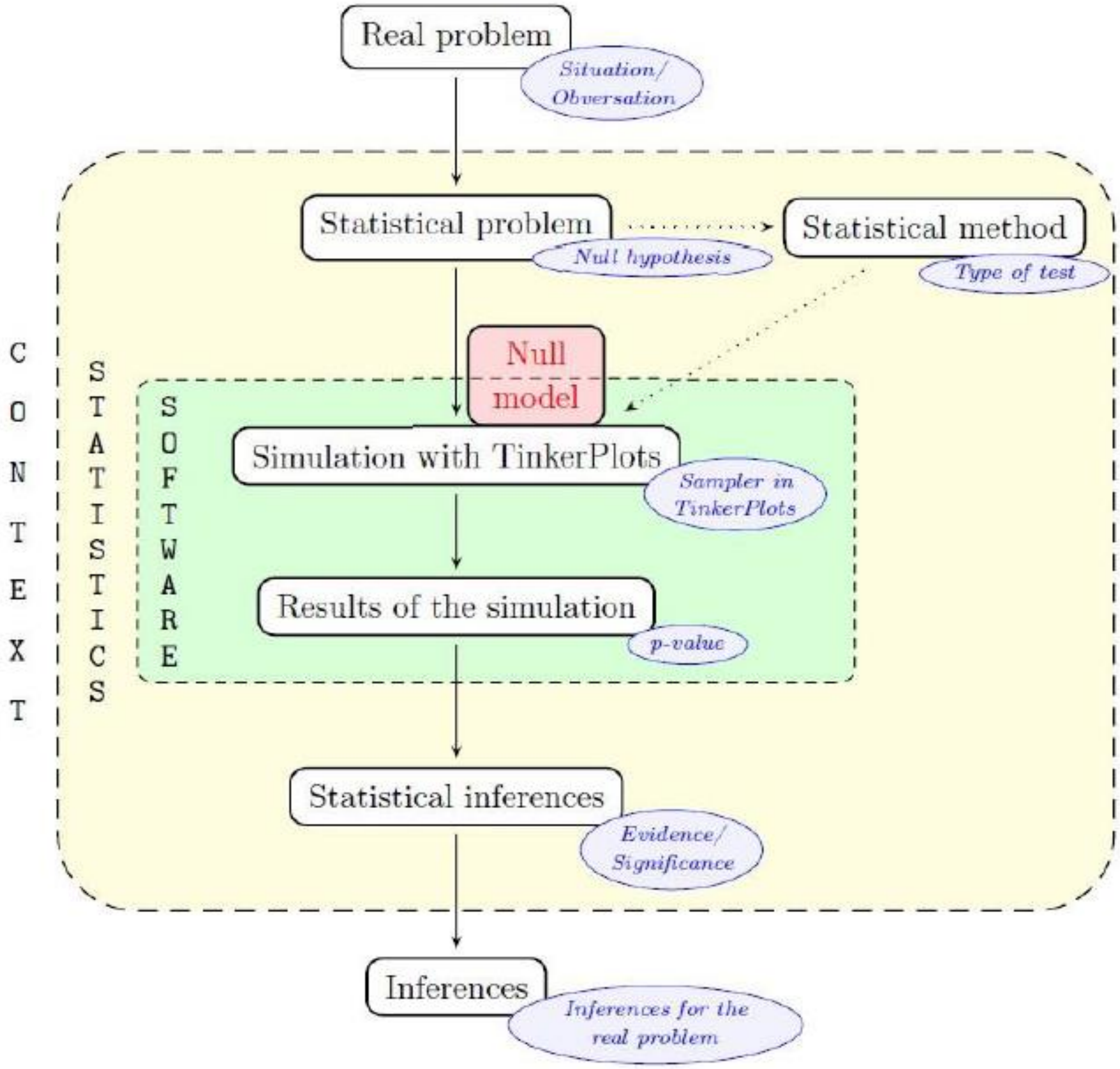
Beyond methodological and philosophical critiques of this method, students have historically struggled to conduct this significance testing procedure. However, new curricula based on simulation appear to lead to larger gains in students’ understanding about significance testing than their mathematical-based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

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

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

Figure 6

Framework for randomization testing (Biehler et al., 2015)



It is perhaps obvious but worthwhile to remember that many students are novices with regards to formally reasoning about probability. When considering the Facebook task (see Figure 3), the Music Note task (see Figure 4), and the NFL task (see Figure 5), a trained probabilist would see the same mechanism in each task – e.g., a simple urn model. This translates to a single TinkerPlots device with two outcomes, the event of interest and its complement.

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

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

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

Another possible explanation for students' struggles through the lens of Biehler et al.'s (2015) framework is that students appear to view the purpose of a null model simulator only in terms of its product (i.e., the results of the simulation), and not for its role as the representation of the null hypothesis and null model (e.g., Justice et al., 2018). While students are generally able to reason 'down' the levels from the results of the simulation to statistical inferences and conclusions about the problem context, they struggle to reason 'up' the level from a real problem to a statistical problem and to the null model simulator (e.g., Noll et al., 2016). Biehler et al. (2015) postulated that some students may have an internal schema for randomization test procedures using TinkerPlots, but lack a conceptual understanding linking a null model simulator, in terms of both its represented process and its enacted product, with statistical inference and the logic of significance testing. While this theoretical framework specifies a 'statistical' level, this may not be a distinct level in students' minds, which may also explain their conflation of process and product when considering the purpose of a null model simulator.

4.1. LIMITATIONS OF CURRENT RESEARCH

As previously discussed, and perhaps by coincidence, most of the studies documenting students' performance on assessment items related to significance tests were evaluations of the ISI curriculum, while most of the studies documenting students' reasoning about significance tests through observation or written work utilized the TinkerPlots software and the CATALST curriculum. While Hildreth et al. (2018) found that students' understanding of statistical inference was comparable between the CATALST, ISI, and Lock5 curricula, TinkerPlots requires its users to create null model simulators from scratch, as opposed to the Rossman-

Chance applets or StatKey which provide pre-constructed null model simulators or require only partial specification of its characteristics. Although these software tools do not require students to explicitly construct null models, it does not mean they do not build an understanding of some or all aspects of the null model, as they still see and interact with the product of the enacted simulation, i.e., the sampling distribution under the null model.

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

4.2. RECOMMENDATIONS FOR FUTURE RESEARCH

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

A focus on the relationship between study design and null models, meant to foster students' understanding of the data generating process that the null model simulator represents, also provides a unique opportunity to incorporate null models in Model Eliciting Activities (MEA; e.g., Garfield et al., 2012) which are already a part of some simulation-based curricula. While MEAs typically ask students to build a model based on real-world patterns, a Null Model Eliciting Activity might instead focus on predicting real-world patterns based on a model. These Null Model Eliciting Activities may aid in the development of students' understanding of the variability specified in null hypotheses and ultimately the transcription of a null hypothesis into a null model simulator.

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

Related to students’ narrative sense-making, it is currently unknown to what extent students’ creation of null model simulators is sensitive to problem task text characteristics and story structure or schema. Through repeated exposure, individuals form a story schema, which is an organizational structure in which they expect information to be provided (Mandler, 1984). Misalignment between a story’s plot and an individual’s story schema can impede processing and storing of information (Anderson & Pearson, 1984; Bartlett, 1932). While colloquial discourse is verbal, dialogical, and often takes the form of a narrative, scientific discourse whether in professional writing or the classroom is monological and relies on definitions, direct descriptions, and explanation in a non-narrative manner (Klein, 2006). Thus, unfamiliarity with the format in which statistical information is presented may create difficulties for students to understand statistical concepts and procedures and similarly impede students in translating a real-world problem into a statistical problem (e.g., Kendeou & Van Den Broek, 2007; Moravcsik & Kintsch, 1993).

Students’ inclusion of narrative characteristics such as temporal sequence when constructing simulators may simply reflect the organizational structure through which they process the problem context via narrative sense-making. This also provides an opportunity for instructors and researchers to rethink the way they construct text presenting statistical tasks to students or explaining statistical information. Future research can explore students’ sensitivities to these characteristics and attempt to design new activities and formative assessments to support students’ statistical information processing.

Finally, the simulation-based approach originally grew out of a desire to cut free from the technical complexity of mathematical-based methods (Cobb, 2007). While a simulated sampling distribution may provide students a more concrete connection to the random process specified under the null hypothesis, it is worth considering whether these reforms push common 20th century methods far enough. Some researchers have decried the entire hypothetico-deductive philosophy of confirmation underlying significance tests for its *malaise* and unsuitability to scientific experimentation (e.g., Rozeboom, 1997). Hypothetico-deductive confirmation is just one of three philosophical approaches to confirmation that could serve as a basis for statistical tests. While Bayesian ideas and methods have begun to proliferate through the introductory level (e.g., Hoegh, 2020), they often carry a similar mathematical complexity that led to calls for reform towards the end of the 20th century.

The likelihood school of statistics, based on the likelihood principle which states that all the relevant information about a sample in terms of its function as evidence in relation to potential hypothetical parameters is contained in the likelihood function (Hacking, 1965), may provide a viable alternative. A test of a hypothesis begins with reasoning about the likelihood function and which hypotheses it supports, akin to Hempel-confirmation. A strong conceptual understanding of likelihood functions could provide a basis for students to further their studies in either the classical school or the Bayesian school, as the classical school utilizes likelihood in maximum likelihood estimation and the Bayesian school combines likelihood functions with prior probability distributions to form posterior distributions. Given students difficulties with the conditional reasoning required by classical null hypotheses, simulation-based likelihood functions may both provide an opportunity to edify their understanding and present an alluring alternative solution to a decades-long debate concerning the utility and validity of these methods. While simulation-based methods may eventually replace mathematical-based method in introductory level statistics instruction, identifying pedagogical and technological tools to support the development of students’ understanding of statistical inference remains paramount.

5. REFERENCES

Anderson, R. C., & Pearson, P. D. (1984). A schemata-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (Vol. 1, pp. 255–291). New York: Longman.

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

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

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

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, MA: MIT Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Cobb, P., & McClain, K. (2004). Principles of Instructional Design for Supporting the Development of Students' Statistical Reasoning. In D. Ben-Zvi & J. Garfield (Eds.) *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking.* Dordrecht: Kluwer Academic Publishers, pp. 375–395.

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

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

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

de Finetti, B. (1980). Foresight: Its logical laws, its subjective sources (H.E. Kyburg Jr., trans.). In H. E. Kyburg, Jr. and H. E. Smokler (Eds.), *Studies in Subjective Probability*. John Wiley and Sons. (Original work published 1937) delMas, R. C., Garfield, J. B., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hoegh, A. (2020). Why Bayesian ideas should be introduced in the statistics curricula and how to do so. *Journal of Statistics Education, 28*(3), 222-228.

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

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

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

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

Kendeou, P., & Van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & cognition, 35*(7), 1567-1577.

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

Klein, P. D. (2006). The challenges of scientific literacy: From the viewpoint of second-generation cognitive science. *International Journal of Science Education, 28*(2–3): 143–178.

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

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

Lindley, D. V. (1975). The future of statistics: a Bayesian 21st century. *Advances in Applied Probability, 7*, 106-115.

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

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

Mandler, J. M. (1984). *Stories, scripts, and scenes: Aspects of schemata theory*. Hillsdale, NJ: Lawrence Erlbaum.

Maxara, C., & Biehler, R. (2007). Constructing stochastic simulations with a computer tool – students' competencies and difficulties. In D. Pitta & P. G. Philippou (Eds.), *Proceedings of the Fifth Conference of the European Society for Research in Mathematics Education* (p. 762-771). Lamaca, Cyprus.

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

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

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

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

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

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

Moravcsik, J. E., & Kintsch, W. (1993). Writing quality, reading skills, and domain knowledge as factors in text comprehension. *Canadian Journal of Experimental Psychology*, 47(2), 360-374.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., … 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., … VanderStoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education*. Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

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

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

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

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

Vallecillos, A. & Batanero, C. (1997). Conceptos activados en el contraste de hipotesis estadısticas y su comprension por estudiantes universitarios [University students' difficulties in understanding key concepts of hypotheses testing]. *Recherches en Didactique des Mathematiques, 17,* 29–48.

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

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

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

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

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

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

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

Zieffler, A., & Catalysts for Change. (2019). *Statistical Thinking: A simulation approach to uncertainty* (4.2th ed.). Minneapolis, MN: Catalyst Press. http://zief0002.github.io/statistical-thinking/

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

Appendix A

Tests of Significance items from CAOS (delMas et al., 2007)

Item No.	Item Stem	Response Options
19	A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain?	a. A large *p*-value. b. A small *p*-value c. The magnitude of a *p*-value has no impact on statistical significance
23	A researcher in environmental science is conducting a study to investigate the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either a treatment or a control group. The first in the treatment group showed higher levels of the indicator enzyme. Suppose a test of significance was correctly conducted and showed no statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?	a. The researcher must not be interpreting the results correctly; there should be a significant difference. b. The sample size may be too small to detect a statistically significant difference. c. It must be true that the herbicide does not cause higher levels of the enzyme.
	A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with Macular Degeneration. The article gives a *p*-value of 0.04 in the analysis section. Items 25, 26, and 27 present three different interpretations of this *p*-value. Indicate if each interpretation is valid or invalid.	
25	The probability of getting results as extreme as or more extreme than the ones in this study if the drug is actually not effective.	a. Valid b. Invalid.
26	The probability that the drug is not effective.	a. Valid. b. Invalid.
27	The probability that the drug is effective.	a. Valid. b. Invalid.
40	The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternate hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?	a. The circuit is definitely not good and needs to be repaired. b. The electrician decides that the circuit is defective, but it could be good. c. The circuit is definitely good and does not need to be repaired. d. The circuit is most likely good, but it could be defective.

Appendix B

Simulation-based Inference topic items from GOALS-4 (Sabbag, 2016; Sabbag et al., 2015)

<table>
<tr><th>Item No.</th><th>Item Stem and Response Options^</th><th>Percent correct*</th></tr>
<tr><td>6</td><td>A researcher investigated the impact of a particular herbicide on the enzyme level of carbonyl reductase in fish. In the study, 60 farm-raised fish were randomly assigned to the treatment group (in which they were exposed to the herbicide) or to the control group (in which they were <i>not</i> exposed to the herbicide). There were 30 fish assigned to each group. After the study, the data were analyzed, and the results of that analysis are reported in the output below. (<i>p</i> = 0.3644; 95% CI: -11.15 – 4.16)

Based on the results of the study, the researchers should not conclude that the herbicide has an effect on the enzyme levels of farm-raised fish.
a. Valid
b. Invalid</td><td>68.3%</td></tr>
<tr><td>14</td><td>Two medical researchers each perform the same experiment using two different samples from the same population. One study results in a <i>p</i>-value of 0.06, and the other study results in a <i>p</i>-value of 0.09. Which of the following statements is correct regarding the evidence against the null hypothesis?
a. The <i>p</i>-value of 0.06 gives stronger evidence against the null hypothesis because it is smaller.
b. The <i>p</i>-value of 0.09 gives stronger evidence against the null hypothesis because it is larger.
c. It’s impossible to tell which <i>p</i>-value provides stronger evidence against the null hypothesis, because they are both greater than 0.05.</td><td>45.2%</td></tr>
<tr><td></td><td>Yolanda was interested in whether offering people financial incentives can improve their performance playing video games. Yolanda designed a study to examine whether video game players are more likely to win a game when they receive a $5 incentive or when they simply receive verbal encouragement. Forty subjects were randomly assigned to one of two groups. The first group was told they would receive $5 if they won the game and the second group received verbal encouragement to “do your best” on the game. Yolanda collected the following data from her study:
<table>
<tr><th></th><th>$5 Incentive</th><th>Verbal Encouragement</th></tr>
<tr><td>Win</td><td>16</td><td>8</td></tr>
<tr><td>Lose</td><td>4</td><td>12</td></tr>
</table>
Based on these data, it appears that the $5 incentive was more successful in improving performance than the verbal encouragement, because the observed difference in the proportion of players who won was (16/20) – (8/20) = 0.40. In order to test whether this observed difference is only due to chance, Yolanda does the following:
- She gets 40 index cards. On 24 she writes, “win” and on 16 she writes, “lose”.
- She then shuffles the cards and randomly places the cards into two stacks of 20 cards each. One stack represents the participants assigned to the $5 incentive group and the other represents the participants assigned to the verbal encouragement group.
- She computes the difference in performance for these two hypothetical groups by subtracting the proportion of winning players in the “verbal encouragement” stack from the proportion of winning players in the “$5 incentive stack”. She records the computed difference on a plot.
- Yolanda repeats the previous three steps 100 times.</td><td></td></tr>
</table>

15	What is the explanation for the process Yolanda followed? a. This process allows her to determine the percentage of time the $5 incentive group would outperform the verbal encouragement group if the experiment were repeated many times. b. This process allows her to determine how many times she needs to replicate the experiment for valid results. c. This process allows her to see how different the two groups' performance would be if both types of incentive were equally effective.	34.0%
16	Yolanda simulated data under which of the following assumptions? a. Verbal encouragement is more effective than a $5 incentive for improving performance. b. The $5 incentive is more effective than verbal encouragement for improving performance. c. The $5 incentive and verbal encouragement are equally effective at improving performance	31.1%
17	Below is a plot of the simulated differences in proportion of wins that Yolanda generated from her 100 trials. Based on this plot, the one-sided *p*-value is 0.03. Which of the following conclusions about the effectiveness of the $5 incentive is valid based on these simulation results? a. The $5 incentive is more effective than verbal encouragement because the *p*-value is less than 0.05. b. The $5 incentive is more effective than verbal encouragement because distribution is centered at 0. c. The $5 incentive is not more effective than verbal encouragement because distribution is centered at 0. d. The $5 incentive is not more effective than verbal encouragement because the *p*-value is less than 0.05.	48.1%
18	The *p*-value is the probability that the $5 incentive group would win more often than the verbal encouragement group. a. Valid. b. Invalid.	41.9%
20	In Yolanda's experiment, there were 20 subjects randomly assigned to each group. Imagine a new study where 100 students were randomly assigned to each of the two groups. Assume that the observed difference in this new study was again 0.40 (i.e., that the proportion of wins for the $5 incentive group was 0.40 higher than the observed proportion of wins for the verbal encouragement group). How would the *p*-value for this new study (100 per group) compare to the *p*-value for the original study (20 per group)? a. It would be the same as the original *p*-value. b. It would be smaller than the original *p*-value. c. It would be larger than the original *p*-value.	44.9%

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

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