# Eliciting and Deciphering Mathematics Teachers' Knowledge in Statistical Thinking, Statistical Teaching, and Statistical Technology

Yu Gu

Submitted in partial fulfillment of the
requirements for the degree of Doctor of Education in
Teachers College, Columbia University

2021

**Abstract**

Eliciting and Deciphering Mathematics Teachers' Knowledge in

Statistical Thinking, Statistical Teaching, and Statistical Technology

Yu Gu

Statistically skilled workers are highly demanded in today's world, which means we need high-quality statistics education. There has been a continuously increased enrollment of statistics students. At the college level, introductory statistics courses are typically taught by professors who often hold a strong qualification in mathematics but may lack formal training in statistics education and statistical analysis. Existing literature claims that a unique way of thinking–statistical thinking or reasoning–is essential when teaching statistics, especially at the introductory level. To elaborate and expand on the issue of statistical thinking, a qualitative study was conducted on 15 mathematics teachers from a local community college to discuss differences between statistics and mathematics as academic disciplines and exemplify two types of thinking–statistical thinking and mathematical thinking–among mathematics teachers who teach college-level introductory statistics. Additionally, the study also inspected mathematics teachers' pedagogical ideas influenced by each type of thinking, some of which were recognized as "pedagogically powerful ideas" that transcend students' conceptual understanding about statistics.

The study consisted of two online questionnaires and one interview. In the two online questionnaires, participants explored and rated five technology options for teaching statistics and self-evaluated their technology, pedagogy, and content knowledge. During the interview, participants solved nine statistical problems designed to elicit statistical thinking and addressed

pertinent pedagogical questions related to each problem's statistical concept. A framework that hypothesizes aspects of mathematics teachers' statistical thinking and mathematical thinking in statistics was created, summarizing the prominent differences in problem-solving, variability, context, data production, transnumeration, and probabilistic thinking. Select responses from participating mathematics teachers were provided as examples of each type of thinking. Furthermore, it was revealed that mathematics teachers with a different type of thinking tended to cover different statistical topics, deliver the same statistical concept in different ways, and assess students' knowledge with different emphases and standards. This study's results have implications: if statistics is to be taught by mathematics teachers, statistical thinking is required to implement pedagogically powerful ideas for furthering meaningful statistical learning and to unveil the differences between statistics and mathematics.

## Table of Contents

## List of Tables

## List of Figures

**Acknowledgments**

Since I started preparing the first draft of this dissertation, I have always imagined the day when I would be listening to melancholy music and writing the acknowledgment page, knowing that I am done with the rest of the dissertation. But this day did not come as smoothly as I naively hoped. As I went further into this journey, I became quite acquainted with the taste of despair. It gradually destroyed my confidence and enthusiasm. And I thought about giving up once for all. Luckily, I was able to share these negative thoughts and emotions with people in my life to whom I forever indebted.

I would like to thank my family and friends, both overseas and in the states, for tolerating my capricious temperament sometimes and convincing me that I can finish my degree. Remarkably, I can't be more grateful for the births of my daughter and son. I love you both to the moon and back.

I would like to thank Dr. Michael Kazlow for enabling me to explore teaching practices with new ideas. I greatly appreciate your supervision and guidance in my teaching endeavor for inspiring the design of this study.

I would like to thank my dissertation committee members. Thank you, Dr. J. Philip Smith, for your continuous support as my dissertation sponsor throughout the entire journey. Especially, thank you for your valuable mentorship and your emotional comfort in my darkest days at Teachers College prior to the approval of the prospectus. Without you, I would not be able to write this page today to mark the end of this journey. Thank you, Dr. Chaya Flint, for your professional advice and genuine encouragement. I am so glad that I took your summer class during my darkest days at Teachers College. Your personality brought sunshine to my

once-very-dark world and gave me the hope that eventually saved me from the abyss of depression and anxiety. Thank you, Dr. Felicia Mensah, for your valuable input on the methodology and your assistant with the organization of the remote defense.

Lastly, I would like to thank all the 15 mathematics teachers whose names remain anonymous. You devoted your precious time to taking part in this very tedious study and helped shape a better future for statistics education. You all deserve a thundering ovation!

Y. G.

**Dedication**

To my mom and dad,

who brought me to this world and took care of me with endless and unconditional love.

To the love of my life,

who accompanied me on my entire dissertation journey and believed in me when I didn't.

To my precious daughter and son,

who taught me what it means and what it takes to be a good father.

I am forever grateful for having all of you in my life.

Chapter 1

# INTRODUCTION

## Need for the Study

The advent of a faster central processing unit in conjunction with the ever-growing storage capabilities in computer science enabled repeated analysis of a massive amount of data within a blink of an eye. Nowadays, a smartphone in an average person's pocket might as well equip with more computing power than all of NASA's computers combined for the evolutionary moon landing mission in 1969 (Puiu, 2019). Why does this matter? Well, theoretically speaking, anyone who's playing match-three games on the subway is holding a device that is powerful enough to perform the computation necessary for sending a human being to the moon! The real question is, is the device holder or the match-three puzzle solver knowledgeable enough to utilize the device held entirely?

In this new era of "big data," technology is making people's life more comfortable than ever but is also secretly making job recruitment much harder. Based on a 2016 survey conducted by the Society for Human Resource Management and sponsored by the American Statistical Association, 78% of organizations reported having difficulty recruiting moderately skilled data analyst, 89% expect their moderate candidates to be able to use specialized software or analysis tools to perform advanced statistical analysis, and 86% expect their moderate candidates to be able to interpret results to other non-statistical background coworkers (the Society for Human Resource Management (SHRM), 2016). Some more recent statistics posted online by the U.S. Department of Labor in 2019 suggest that the job market for people

with statistical reasoning skills is estimated to increase by 34% from 2016 to 2026, exceeding the average of all occupations (Bureau of Labor Statistics, 2019). We need more data analysts. In other words, we need sufficiently enough statistics education.

In fact, statistics education has always been placed at an increasing priority in the history of the American school curriculum. In 1923, triggered by the rapid development of industrialization, the Mathematical Association of America (MAA) recommended the inclusion of elementary statistics in the junior high and high school curriculum. In 1931, Helen Walker, a statistics professor at Teachers College, Columbia University, became the first statistician to openly support the instruction of statistics to young boys and girls. During World War II, the National Research Council's (NRC) Committee on Applied Mathematical Statistics realized the importance of statistics in secondary education but admitted the shortage of qualified statistics teachers. While competing with Russia in space war and advanced weapon technology, Frederick Mosteller, the president of the American Statistical Association (ASA) back then, established a joint committee with the National Council of Teachers of Mathematics (NCTM) on school curriculum development in statistics. The high school curriculum in statistics became more real-life-based and data-analysis-oriented. In 1980, NCTM published recommendations on new statistical topics that should be included in the school mathematics curriculum, such as collection, summarization, presentation, interpretation, and data prediction. Not long after, with funding granted from the National Science Foundation (NSF), the ASA and NCTM Joint Committee launched a Quantitative Literacy Project (QLP), which developed rich materials for statistical teaching and inspired NCTM for publishing the *Curriculum and Evaluation Standards for School Mathematics* in 1989. Statistics became an integral part of the mathematics curriculum ever since and gradually led to the debut and success of Advanced Placement (AP)

Statistics courses in high schools taught by leaders in QLP. A revolutionary turning point occurred in 2007 when the *Guidelines for Assessment and Instruction in Statistics Education: A PreK–12 Curriculum Framework (GAISE)* was released to the public. This GAISE report brought up the significance of differentiating thinkings between mathematics and statistics and proposed a theoretical framework for assessing learner's developmental level of statistical understanding. Three years later, many states in the U.S. adopted the Common Core State Standards for Mathematics (CCSSM), which was created based on the GAISE framework (Franklin et al., 2015).

Perhaps the most prominent consequence of this continuous growth of statistics in the school curriculum is the increased enrollment of students learning statistics at secondary and tertiary levels.

> According to the Conference Board on Mathematical Sciences (CBMS) survey, 508,000 students took an introductory statistics course in a two- or four-year college/university in the fall of 2010, a 34.7% increase from 2005. More than a quarter (27.0%) of these enrollments were at two-year colleges. Nearly 200,000 students took the Advanced Placement (AP) Statistics exam in 2015, an increase of more than 150% over 2005. In addition, many high school students took the AP course without taking the exam or took a non-AP statistics course. At the undergraduate level, the number of students completing an undergraduate major in Statistics grew by more than 140% between 2003 and 2013 and continues to grow rapidly. (Carver et al., 2016, p. 4)

But why do we still lack people with data analysis skills? Quantity does not guarantee quality. Maybe our students are not receiving statistics education that cultivates essential statistical analysis skills. Could the shortage of qualified statistics teachers still exist as it has been observed since the 1940s? After an extensive literature review pertaining to studies on teachers' statistical knowledge, Shaughnessy (2007) concluded that most teachers teaching introductory statistics courses are labeled "undernourished" in statistical knowledge due to the

deficiency of their own statistics education when they were students. But there's more. Even those who demonstrated a strong mathematical background are not necessarily qualified to teach statistics. Hannigan, Gill, and Leavy (2013) applied the Comprehensive Assessment of Outcomes in a First Statistics (CAOS) course test to investigate pre-service mathematics teachers' both mathematical and statistical knowledge. Results showed that pre-service teachers with science major background did much better in mathematics. But surprisingly, their statistics results were poor and no better than those obtained by their non-science background classmates. To complement this quantitative result, a qualitative study conducted by Leavy, Hannigan, and Fitzmaurice (2013) concluded that mathematics teachers did find the unique statistical thinking and reasoning difficult to grasp. There exist an inequality and an intransitivity between mathematical thinking and statistical thinking.

> Across all levels and stages of the investigative process, statistics anticipates and accounts for variability in data. Whereas mathematics answers deterministic questions, statistics provides a coherent set of tools for dealing with . . . natural variability in populations, induced variability in experiments, and sampling variability in a statistic, to name a few. The focus on variability distinguishes statistical content from mathematical content. For example, designing studies that control variability, using distributions to describe variability, and drawing inferences about a population based on a sample in light of sampling variability all require content knowledge distinct from mathematics. In addition to these differences in content, statistical reasoning is distinct from mathematical reasoning, as the former is inextricably linked to context. Reasoning in mathematics leads to the discovery of mathematical patterns underlying the context. In contrast, statistical reasoning is necessarily dependent on data and context and requires an integration of concrete and abstract ideas. (Franklin et al., 2015, p. 1)

In summary, the interpretations of numbers in mathematics are theoretically computable and predictable without the necessity of context. In contrast, interpreting results in statistics is bound by context, variability, and uncertainty (Franklin et al., 2007). Mathematical

thinking tends to rely on formal structures and rigorous procedures to seek a theoretical solution, whereas statistical thinking tends toward a more fluid, tentative analysis to make practical inferences about a long-term behavior. Irrefutably, an appreciation of such observations from introductory statistics instructors will have invaluable pedagogical implications. But a more systematically defined description of the differences between statistical thinking and mathematical thinking is needed to increase the awareness of teaching statistics in its "intended" way. As a result, statistics students will have a chance to view statistics differently from mathematics and imbibe the statistical knowledge in a meaningful way.

Unfortunately, the current situation is not optimistic. Studies have shown that most pre-service teachers cannot observe the differences between statistics and mathematics, which could negatively impact their future students' statistical thinking (Hannigan et al., 2013; Leavy et al., 2013; Shaughnessy, 2007). However, not much has been reported on in-service or college mathematics teachers' statistical thinking. Are they able to sense the difference? Are they able to take the great responsibility to educate our match-three puzzle solvers to become moderate data analysts? This study attempted to address such inquiries.

Another worth-mentioning consequence of the growing attention in statistics education is the emergence of a diverse list of statistical technology options for both statistical research and teaching.

> These include course management systems, automated homework systems, technology for facilitating discussion and engagement, audience response systems, and videos now used in many courses. Applets and other applications, such as Shiny apps coded in the R programming language, that are designed to explore statistical concepts have come into widespread use. Many general-purpose

> statistical packages have developed functions specifically for teaching and learning. (Carver et al., 2016, p. 5)

Course management systems provide new ways of communication between teachers and students as well as online collaboration between students (Ben-Zvi, 2007). Easily accessible web-applets allow teachers to present difficult introductory statistical concepts, such as confidence levels, sampling distributions, and $p$-values, in a more intuitive representation through the technique of "resampling statistics," which, in the old days, required knowledge of specific statistical packages and computer programming from teachers (Good, 2006). The prevalence of dynamic statistical software such as Tinkerplots has been demonstrated to promote "what-if" moments and furthering statistical understanding (Ben-Zvi, Aridor, Makar, & Bakker, 2012; Lehrer, Kim, & Schauble, 2007; Watson & Donne, 2009).

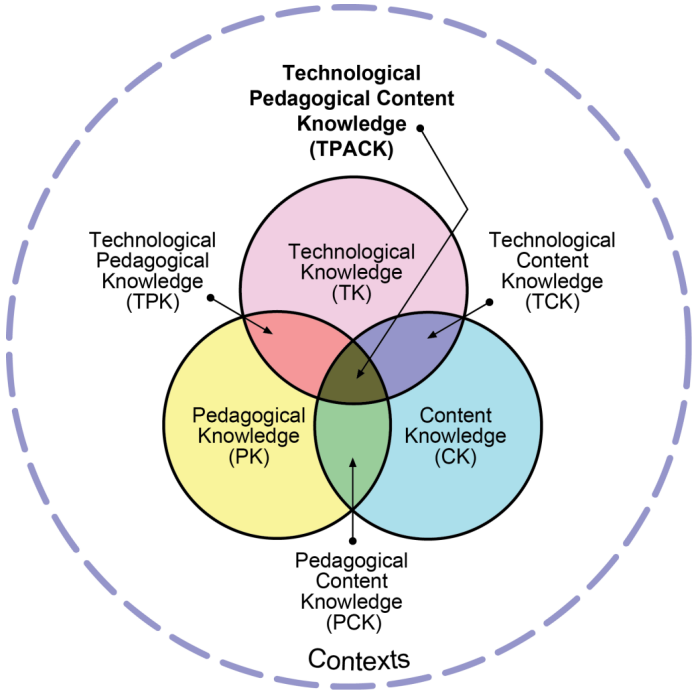
Regardless of numerous reported benefits of technology integration in statistics education, however, using technology does not automatically improve learners' understanding or the quality of education (Zbiek, Heid, Blume, & Dick, 2007). Instead of reproducing existing teaching instruments, the role of technology in statistics education should transform students' learning experience of statistics from rote computation to exploratory data analysis and heterogeneous data inference (Hooper & Rieber, 1995).

## Purpose of the Study

Contemporary statistical technology is not yet advanced enough to replace teachers. Students need guidance from teachers in choosing and using statistical technology (Forster, 2006). That requires teachers to possess knowledge other than statistical content knowledge. According to Koehler and Mishra (2009), teachers who wish to adopt technology in the classroom successfully demand knowledge in three intersecting domains: content, pedagogy,

and technology. They proposed a theoretical framework, called the *technological pedagogical content knowledge* (TPACK) that contains seven knowledge subdomains (see Figure 1.1). It is an extension of the *Pedagogical Content Knowledge* (PCK) framework created by Shulman (1986, 1987). As the prevalence of technology integration in education permeates, many recent studies have adopted the TPACK framework to access pre-service teachers' TPACK and their decision-making in technology integration outside a specific discipline domain. Some studies modified the TPACK framework to suit specific discipline's unique context and domain (Chai, Koh, & Tsai, 2016; Willermark, 2018). In statistics education, for instance, Lee and Hollebrands (2011) developed the *Technological Pedagogical Statistical Knowledge* (TPSK) framework for evaluating teachers' TPACK in statistics. But, so far, only a few studies have been reported (Birel & Çakıroğlu, 2018; Henriques & Gutiérrez-Fallas, 2017). It was the researcher's hope that

Figure 1.1. *The TPACK framework and Its Seven Knowledge Subdomains. Reproduced by permission of the publisher, ©2019 by tpack.org.*



this study attempted to address the lack of research on college teachers and the assessment of the TPACK in statistics.

The purpose of the study is to elicit and decipher college mathematics teachers' knowledge in statistical thinking, statistical teaching, and statistical technology at the college introductory statistics level. Particularly, this study attempts to 1) investigate the differences between mathematical thinking and statistical thinking when mathematics teachers understand various introductory statistical concepts; 2) explore connections between these understandings and teachers' pedagogical choices on teaching topics in descriptive statistics and inferential statistics with or without technology; 3) identify key developmental understandings and pedagogically powerful ideas from mathematics teachers that promote meaningful statistical learning.

Explicitly, the study serves to answer the following research questions.

1. In which ways do statistical thinking and mathematical thinking take place among mathematics teachers when teaching introductory statistics?
2. With a general knowledge of various types of statistical technology options, how does mathematics teachers' statistical thinking or mathematical thinking affect their way of teaching?
3. With a general knowledge of various types of statistical technology options, how do mathematics teachers promote statistical learning through teaching with or without technology?

## Procedures of the Study

An array of methods have been reported and used to assess teachers' TPACK and their decision-making in technology integration, including but not limited to survey, open-ended questionnaires, concept mapping, lesson plan analysis, interviews, performance assessments, and observations (Rosenberg, 2012). In this study, a combination of online surveys and in-person interviews was adopted, providing qualitative data from teachers to understand better their TPACK in statistics and their primary type of thinking when addressing different topics in a typical introductory statistics course.

### Sampling

Potential candidates of this study were selected from mathematics teachers who were teaching or had taught college-level introductory statistics in New York City. Diverse academic backgrounds from the candidates were identified with different academic achievements and teaching experiences. Even though 19 teachers initiated their participation in the study, only 15 teachers completed the entire process. A two-letter code was randomly generated and assigned to each participant upon recruitment as identification.

### Data Collection

*Stage 1: Before the Interview*

**Phase I**: Technology Evaluation. Each participant received a link to an online survey (see Appendix A) regarding the evaluation of five technology options. The survey collected demographic information and teaching experience from participants. It also simulated a scenario where teachers tried to learn about potential statistical technology for their teaching

in real-life. By going through the survey, participants gained an overview of five types of technology options available for teaching statistics:

1. Statistical Software Package: computer programs created for professional statistical analysis.
2. Educational Software: computer programs created for teaching statistics.
3. Spreadsheets: Excel or Google Sheets style programs widely used in finance and business.
4. Web Applets: web applications that typically run in a web browser and allow users to manipulate data in a visual, dynamic, and exploratory manner.
5. Multimedia Materials: computer programs or websites that combine text, audio, images, animation, video, quizzes, and other interactive media in introductory statistics.

Participants were free to explore each option as long as they wanted. Once the exploration was done, they rated each technology option and provided comments on the pros and cons they observed. However, upon completing this survey, participants were not required to know how to use these options.

**Phase II**: Statistical TPACK survey (see Appendix B). Once phase I was complete, a second link was sent out to the participants. This new online questionnaire assessed participants' quantified scores in each domain of the TPACK. The survey was adapted from an existing one made by the creators of the TPACK framework (Schmidt et al., 2009), specifically tailored by the researcher for introductory statistics.

*Stage 2: On the Interview Day*

Each interview lasted for approximately 90 minutes and was audio-taped. Each participant was provided with a tablet and a stylus during the in-person interview. The interview handout was distributed electronically via the tablet at the beginning of the interview. A note-taking application called *Notability* was used to record the interview audio and participant's handwriting on the tablet. For online interviews, the video conference application *Zoom* was used. The electronic version of the handout was screen-shared with the participant. The entire session was recorded (with the participant's camera off), and, if applicable, all scratch work made by the participant was collected electronically via pictures taken by each participant.

**Phase III**: Technology Discussion. Participants went over their technology evaluation from phase I with the researcher. Each participant filled out a chart with summarized pros and cons of each technology type as a reference for the next phase.

**Phase IV**: Statistical Thinking Assessment (see Appendix C). During this core part of the interview session, each participant was shown a series of statistical problems with additional questions related to pedagogy. Topics covered both descriptive statistics and inferential statistics. Problems were selected from the LOCUS[1] assessment which emphasizes on four essential steps of statistical problem solving recommended by the GAISE report (Franklin et al., 2007). The pedagogical questions were created to provoke the different thinking between statistics and mathematics. All participants verbally answered each problem and provided rationales, but some wrote down their problem-solving work. Additionally, they

[1]LOCUS is an NSF Funded DRK12 (DRL-1118168) project focused on developing assessments of statistical understanding.

responded to questions that could reveal their thinking process of statistical problem-solving. For the sake of mimicking an actual testing environment, no correct answer was provided, nor participant's response was judged unless requested by the participant.

**Data Analysis**

The first step to data analysis for this study was to create a *study profile* for each participant. This step was done by assembling online survey results and transcribing each interview session. The results from the TPACK survey provided an overview of the TPACK for each participant as well as the entire sample. Each transcribed interview session was carefully coded based on the constructs provided by the following four theoretical frameworks used for this study:

- *Grounded Theory*: a widely used framework to analyze qualitative data and develop categories and themes that well summarize the findings (Corbin & Strauss, 2014).
- *GAISE framework*: a well written guidance on how to teach statistics that promotes statistical thinking at different levels (Franklin et al., 2007). This framework was used as the primary source for identifying the prominent differences between statistical thinking and mathematical thinking during the 4 investigative steps of statistical problem solving: formulating questions, collecting data, analyzing data, and interpreting results.
- *Statistical Knowledge for Teaching (SKT) framework I & II*: two hypothetical frameworks that classify statistical knowledge based on its relation to pedagogy (Groth, 2007, 2013). The SKT framework I complemented the lack of content-specific constructs in the TPACK framework for statistics and provided examples of tasks in introductory statistics that reflected statistical thinking and mathematical thinking for learning and teaching

different statistical concepts. Additionally, the SKT framework II was used to identify pedagogically powerful ideas in statistics that significantly advanced students' conceptual understanding of statistics.

To answer the first research question in relation to how statistical thinking and mathematical thinking take place among mathematics teachers when teaching introductory statistics, responses to statistical problems in phase IV were compared and summarized among all participants. Individual responses from participants were discussed in detail as typical examples of different ways in which mathematics teachers addressed statistical problems. Then, mathematics teachers' ways of solving statistical problems were further labeled as mathematical thinking or statistical thinking based on the SKT framework I (Groth, 2007) and the grading rubrics provided by LOTUS. The differences between mathematical thinking and statistical thinking were presented side by side for each investigative step in statistical problem solving according to the GAISE framework (Franklin et al., 2007).

To respond to the second research question regarding how mathematics teachers' thinking affects their teaching, responses to pedagogical questions in phase IV were analyzed and categorized for each participant and across all participants based on the SKT framework I (Groth, 2007). Mathematics teachers' pedagogical choices were reported for all participants. Some pedagogical choices were substantiated by mathematics teachers' own rationale. To address the effect of mathematics teachers' thinking on teaching, each participant's pedagogical choice, if there was any, was matched with thinking based on the results from research question 1. Possible patterns of pedagogical choices were identified across all participants for each statistical concept addressed in the interview. Effects were summarized in three areas: topic coverage in statistics, delivery methods in class, and student assessment.

To address the third research question in regards to how mathematics teachers promote statistical learning, all responses in phase IV were analyzed for each participant and across all participants. In particular, by referring to the SKT framework II (Groth, 2013) and studies that encourage exploratory and informal analysis in statistics (Bakker, Derry, & Konold, 2006; Cobb & Moore, 1997; Gil & Ben-Zvi, 2011; Konold & Harradine, 2014; Makar, 2014), select mathematics teachers' pedagogical ideas were identified as pedagogically powerful ideas and KDU that could potentially transcend students' learning experiences in introductory statistics into something powerful and meaningful.

# Chapter 2

# LITERATURE REVIEW

It may not be surprising for many people if an introductory statistics course is taught by a mathematics teacher instead of a statistics teacher. Statistics are numbers, and numbers are often involved in mathematical computations. But, is statistics the same as mathematics, though?

In an article titled "Should Mathematicians Teach Statistics," Moore (1988) gave us a very straightforward answer more than 30 years ago.

> No! Statistics is no more a branch of mathematics than is economics, and should no more be taught by mathematicians. It is a separate discipline that makes heavy and essential use of mathematical tools, but has origins, subject matter, foundational questions and standards that are distinct from those of mathematics. It is true that many advanced texts and research papers in statistics use formidable mathematics, but this is misleading. After all, many a graduate microeconomics text cites the Kuhn-Tucker theorem on the first page, and many research papers in physics are intensely mathematical. Statistics is as much a distinct discipline as are economics and physics. Its subject matter is data and inference from data. It is unprofessional for mathematicians who lack training and experience in working with data to teach statistics. (p. 3)

Moore (1988) claimed that statistics is a separate discipline from mathematics. First, statistics is a discipline developed from "official and private data-gathering," not mathematics. Statistics did not branch out from mathematics as a subdomain. Statistics, as a methodological discipline, provides ways for other disciplines to deal with data. Second, though many statistical analysis utilizes mathematical model, the process of drawing inference or "quantifying uncertainty" from data in statistics is heavily influenced by the method of selection, the data of collection,

and the judgement of the researcher. Statistical knowledge is ranked by its "usefulness in the study of data" rather than its "mathematical depth."

> In my opinion, introductory courses that contain mathematically false statements but require students to work with data are less damaging than courses consisting solely of correct proofs of true theorems. (p. 7)

Third, while statistics uses mathematics for its theoretical foundation, many statistical concepts are rarely applied in the field of mathematics. If statistics and mathematics are the same disciplines, there should be "interrelationships" between them. Mathematicians who have obtained a PhD in mathematics but lack training and experience with real-life data analysis do not equip sufficient knowledge to teach statistics. When it comes to delivering the essences of statistics, they do not have the right way of thinking.

## Statistical Thinking and Mathematical Thinking

If mathematicians should not teach statistics because they lack the right way of thinking, what is the right way to think in statistics? In this section, statistical thinking and mathematical thinking are differentiated based on several aspects: the problem-solving process, variability, context, data production, data representation, and probabilistic thinking.

### Problem Solving Process

When it comes to mathematical problem solving, the famous Hungarian mathematician and mathematics educator, Polya (1971), proposed the following four steps:

1. Understand the problem.
2. Devise a plan.
3. Carry out the plan.
4. Look back.

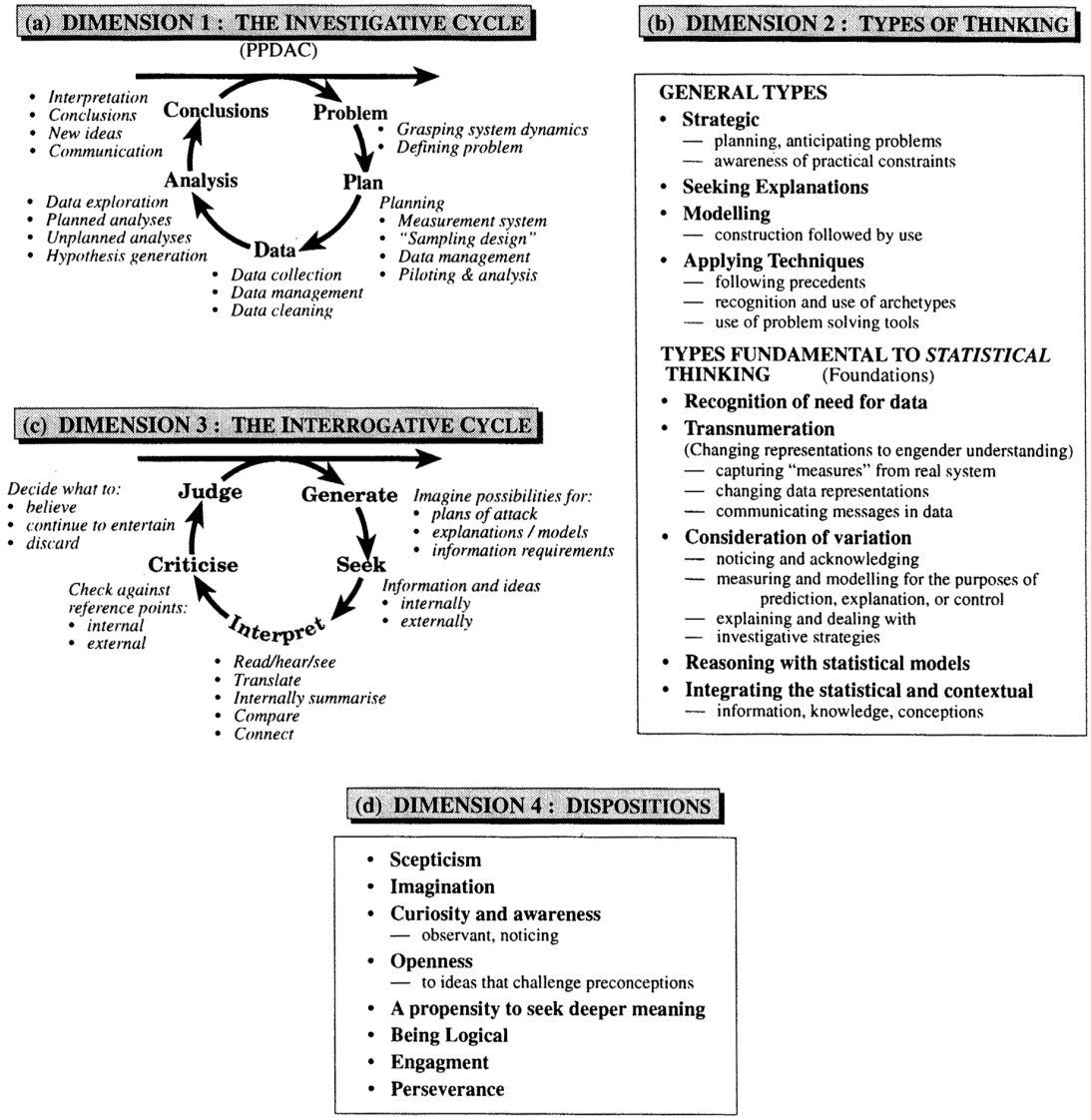
Similarly, Schoenfeld (1981) characterized a typical mathematical problem-solving process into six unordered and recurrent "episodes" – "periods of time during which the problem solver(s) is engaged on a single set of like activities" (p. 4), in order to analyze mathematical learners' behaviors during problem-solving.

1. Read the problem and know its conditions and goals.
2. Analyze the problem and understand its conditions and goals, i.e., simplifying or reformulating the problem.
3. Explore different aspects of the problem, i.e., making connections to certain concepts, past approaches, or similar questions.
4. Plan and implement a select strategy and attempt to solve the problem.
5. Review the problem when failing to solve the problem.
6. Interpret and verify the solution when obtaining an answer.

What about statistical problem-solving? In 1999, based on extensive literature reviews and comprehensive interviews with statistics students and statisticians, Wild and Pfannkuch (1999) proposed a four-dimensional framework for statistical problem solving (see Figure 2.1).

The first dimension consists of the investigative cycle, also called the PPDAC cycle, which iterates among *problem*, *plan*, *data*, *analysis*, and *conclusions*. People conducting statistical analysis often start with contemplating what they want to find, how they want to proceed, and why they decide to do so. The problem step organizes and formulates these vague thoughts into more precise representations in the form of a statistical question that can be answered using data. The planning step identifies and justifies a scientific method to collect data that can be used to address the statistical question. Once the data are collected, reorganized, and cleansed after the data step, they will be analyzed and made sense of during

Figure 2.1. *A 4-Dimensional Framework for Statistical Thinking in Empirical Enquiry. Reprinted from Wild and Pfannkuch, 1999, p. 226.*




the analysis step. The conclusions step conveys and interprets the findings in an attempt to answer the statistical question, which might engender new ideas or information that could amend all previous steps, resulting in moving back and forth within the PPDAC cycle.

The second dimension presents the types of thinking that occur during a statistical problem-solving. They include general thinkings such as strategic, seeking explanations, modeling, and applying mathematical model, as well as those that are fundamental to statistical

thinking: recognition for context, variability, and data production; the ability to transform data into representations better for understanding (*transnumeration*); the knowledge of unique statistical models apposite to the problem situation.

The third dimension depicts the interrogative cycle that reiterates through generate, seek, interpret, criticize, and judge. This cycle is perceived as a sub-cycle for every component of the PPDAC cycle, guiding every step of decision-making.

The fourth dimension summarizes the good characteristics of statistics practitioners that positively initiate and influence statistical thinking: skepticism, imagination, curiosity and awareness, openness, a propensity to seek deeper meaning, being logical, engagement, and perseverance.

Not long after, the guidelines for assessment and instruction in statistics education (GAISE) report (Franklin et al., 2007) stated the ultimate goal of statistics education: *statistical literacy*. According to the report, statistical literacy refers to comprehending and reasoning with daily statistical information around us and drawing statistically sound conclusions or making educated decisions. The GAISE report proposed a conceptual framework for K-12 statistics education to promote a sequential development of statistical literacy among students to reach this goal. The framework broke down the statistical problem-solving into four investigative steps, very similar to Wild and Pfannkuch (1999)'s PPDAC model.

1. Formulate statistical questions that evoke variability in responses.
2. Collect data through a sampling method designed for recognizing and primarily reducing variability in data.
3. Analyze data that accounts for variability in data, i.e., using the appropriate mathematical model for describing the distribution of a population parameter estimate.

4. Interpret results that make generalizations with the inclusion of variability in data.

These four steps provide key components in a typical statistical analysis. Yet, similar to mathematical problem-solving, the actual process of data analysis does not necessarily follow the exact order.

> [F]]or example, choice of design for data production determines the structure of the resulting data, but knowledge based on data already in hand can help shape the design, as when knowing the size of variation from one subject to another helps decide how many subjects will be needed. Similarly, the data may suggest a model, but the model leads to methods that send us back to the data to check for possible violations of the model's assumptions. (Cobb & Moore, 1997, p. 804)

Notably, there are two steps unique to statistical problem-solving: formulate questions and collect data. The current studies suggest that there is a lack of research and teaching practice in these two domains of statistics education (Wild, Utts, & Horton, 2018). As Pfannkuch and Wild (2004) mentioned,

> The foundations of statistical enquiry rest on the assumption that many real situations cannot be judged without the gathering and analysis of properly collected data. Anecdotal evidence or one's own experience may be unreliable and misleading for judgments and decision making. Therefore, properly collected data are considered a prime requirement for reliable judgments about real situations. (p. 18)

When formulating a question in statistical problem-solving, the questions need to be a statistical question that evokes and expects variability in responses. When collecting data, the learners need to know what contributes to a sound statistical design and what the limitations are for the current design (Watson, 2017).

Both mathematical problem-solving and statistical problem-solving include analysis and interpretation. But analysis and interpretation are oriented differently in each case. Mathematics tends to focus on the rigorousness in rules, properties, formulas, and theorems

applied to a particular method or model. In the end, mathematical problem-solving looks for a pattern that theoretically resolves the problem the best (Cobb & Moore, 1997) and addresses the "why" question (Groth, 2013). A genuine statistical analysis includes the mathematical process aforementioned, though it is mostly automated by statistical software and less theoretical at the introductory level (Shaughnessy, 2007). But more importantly, statistics heavily rely on exploratory analysis before the formal inference, which could be descriptive statistics, data visualization, and informal discussion of patterns and striking deviations observed (Cobb & Moore, 1997). Interpretation in statistical analysis is affected by the research questions formulated, the data collected, the methodologies implemented, the mathematical model applied, and the practical significance implied (Groth, 2007). The ultimate goal of statistical problem-solving is to address the long-term behavior (Groth, 2013) and seek a meaningful pattern rather than a best-fitting one as mathematics often desires (Cobb & Moore, 1997; Hand, 1998). As John Tukey, who advocated for exploratory data analysis, once said, "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

**Variability**

While mathematicians seek an answer that is consensual among other mathematicians, statisticians dream about finding, understanding, quantifying, and interpreting data variability.

A key concept arises in statistical problem-solving, as frequently mentioned in literature (Carver et al., 2016; Cobb & Moore, 1997; Franklin et al., 2015; Franklin et al., 2007; Groth, 2007, 2013; Lee & Hollebrands, 2011; Moore, 1988; Shaughnessy, 2007; Wild & Pfannkuch, 1999), is the omnipresence of variability in data – the core understanding in

statistical thinking. The GAISE framework (Franklin et al., 2007) classified different kinds of variability in statistics:

- *Measurement Variability*: the difference among repeated measurements on the same individual, caused by inconsistent measurement system or uncertainty in quantity being measured.
- *Natural Variability*: the difference among individuals being measured, caused by individuals being different in nature.
- *Induced Variability*: the difference among similar individuals being measured (usually via a controlled environment) caused by subjects being exposed to various factors or conditions.
- *Sampling Variability*: the difference in population estimates (sample statistics) among samples of the same size drawn from the same population, caused by unlike subjects contained in those samples.
- *Chance Variability*: the difference in outcomes observed from repeated sampling and experimental design, caused by different sampling processes, and can be described with probability models.

These variabilities are then embedded within three developmental levels of a statistical problem-solving process.

- Level A: At the beginning level, students' statistical problem-solving is limited to one sample, focusing on the measurement variability, natural variability, and induced variability of individuals within the sample.
- Level B: At the intermediate level, students' statistical problem solving is extended to multiple samples, considering the measurement variability, natural variability, and

induced variability within each sample and between each sample, as well as the sampling variability caused by sampling.

- Level C: At the advanced level, students' statistical problem solving exceeds the scope of samples by generalizing results, such as creating a model or conducting a hypothesis testing, and interpreting the variability resulting from chance or randomization. They will address questions such as how well the model fits the data or whether and why the result is statistically significant.

The GAISE framework recommended that statistics education should follow these three developmental levels. Regardless of the grade level, all statistics learners should start from level A and progress gradually to higher levels. At level A, students will follow teacher's guidance. At levels B and C, students will learn to explore and learn statistics and develop a sense of data on their own.

**Context**

Another way to distinguish statistics from mathematics is through context. In mathematics, context usually serves as a motivation to make the problem more relevant to learners (del Mas, 2004). A person who perceives statistics mathematically often reasons abstractly by identifying patterns and concluding generalizations (Hill & Ball, 2004). When analyzing data in statistics, the person removes any distractions such as the backstory, translates words into mathematical symbols, and applies formulas to obtain numerical results. The final results are usually remarked as deterministic and final (Cobb & Moore, 1997). By the essence of statistics, a person who perceives statistics statistically will go a step further by acknowledging the "omnipresence of variability" in data analysis and continually assessing the

rationality and precision of the results (Pfannkuch & Wild, 2004). Their statistical problem-solving process depends on the characteristics of context as it steers the selection and creation of appropriate mathematical models (del Mas, 2004). They seek to find a suitable model to represent the problem and extract meaningful information from the data (Wild & Pfannkuch, 1999). They remain incredulous about the results, which, in their mind, are partially indeterministic within bounds, or conditional conclusions (Groth, 2007). For example, outliers in statistics can be determined by mathematical formulas and rules. However, to tell whether these outliers originate from recording errors or natural variability, context is needed (Franklin et al., 2007). As Cobb and Moore (1997) summarized, "In mathematics, context obscures structure … In data analysis, context provides meaning" (p. 803).

**Data Production**

Related to context, but not tantamount, is how data are produced in statistical problem-solving. The following example illustrates how the same mathematical model can be applied regardless of how the data were collected.

> Does increasing the amount of calcium in our diet reduce blood pressure? The following numbers give the decrease after 12 weeks in systolic blood pressure for 21 human subjects. The 10 subjects in Group 1 took a calcium supplement for 12 weeks; the 11 in Group 2 took a placebo. Test the hypothesis that the calcium had no effect on blood pressure.
>
> Group 1 (calcium): $7, -4, 17, 17, -3, -5, 1, 10, 11, -2$
>
> Group 2 (placebo): $-1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3$ (Cobb & Moore, 1997, p. 804)

This problem may look like a typical exercise in an average introductory statistics textbook. But Cobb and Moore (1997) deemed otherwise. In their opinion, a person who engages in mathematical thinking will probably think it is another routine "mathematical" problem in statistics and carry out the standard computation procedure to draw a conclusion. However, a

person who engages in statistical thinking may ask, "where did the data come from?" The problem did not explicitly state whether the data were collected from a randomized comparative experiment or an observational study such as a voluntary survey. It is questionable whether or not the standard mathematical model for inference concerning two sample means should be applied or executed. And it is troublesome for someone to draw conclusions based on the inferential statistics computed from these numbers without reflecting on the source of the data.

**Transnumeration**

Wild and Pfannkuch (1999) introduced *transnumeration* as one of the types of fundamental statistical thinking in its four-dimensional framework for statistical problem-solving, and defined it as "numeracy transformations made to facilitate understanding" (p. 227). It is used to describe the process of "forming and changing data representations of aspects of a system to arrive at a better understanding of that system" (p. 227), which takes place through all stages of statistical analysis. In the beginning, transnumeration creates useful graphical representations and enables informal exploration of the data (Gigerenzer & Edwards, 2003). When having doubts about the graphical representations and their implications, transnumeration helps clarify the confusion and eliminate deceivable representations (Chick, Pfannkuch, & Watson, 2005; Shaughnessy & Pfannkuch, 2002). For instance, studies have shown that *hat plot*, a plot that combines a dot plot and a box plot (with median removed), helps statistics students connect individual data values to the aggregate view of data distribution (Bakker, Biehler, & Konold, 2004; Watson, Fitzallen, Wilson, & Creed, 2008). Moreover, while analyzing data formally, transnumeration

allows new perceptions of the data and possibly new statistical models through data transformations and reclassifications (Konold & Higgins, 2003). At the end of statistical analysis, transnumeration selects the best data representations to narrate the most "story" about the data (Monk, 2003).

**Probabilistic Thinking: the Intersection**

Probability, one "noteworthy intersection between statistics and mathematics" (Franklin et al., 2015, p. 1), is not only crucial in subfields of mathematics such as applied mathematics and mathematical modeling (Cobb & Moore, 1997), but also essential to understanding inferential statistics (Franklin et al., 2007). In most college-level introductory statistics courses, probability is used as the theoretical foundation for the *frequentist* model which predicts the long-term behavior of a population based on only the data collected from a well-designed experiment under a certain assumption (null hypothesis). The population parameter to be estimated is unknown and the null hypothesis is assumed to be true (Bayarri & Berger, 2004). For instance, probability in introductory statistics is commonly used to interpret fairness, select representative samples, and understand associations between two categorical variables (Franklin et al., 2007). At the advanced level, probability is used to address "What would happen if I repeat this many many times?" (Cobb & Moore, 1997). Probability is heavily used in the confidence interval as the confidence level to quantify the success rate of a particular process capturing the true value of the parameter (del Mas, 2004; Fidler & Loftus, 2009). Additionally, hypothesis testing, or test of significance, also known as the null hypothesis significance testing (NHST), is another probabilistic inference method based on the theoretically infinite samples taken from the population with replacement (Goodman, 2008).

"Probabilistic thinking occurs when reasoning with theoretical probability models, for example, in situations where the argument is based on the data being a random sample from a particular model" (Pfannkuch & Wild, 2004, p. 36). The failure of understanding probability will lead to the belief that a single selected sample reflects all characteristics of the population (Franklin et al., 2015; Franklin et al., 2007; Yu et al., 1995). Pfannkuch et al. (2016) interviewed seven statistics practitioners in probability modeling and characterized their probabilistic thinking as a recursive cycle, called the SWAMTU, that iterates over *problem Situation*, *Want to know*, *modeling Assumptions*, *build stochastic Model*, *Test model*, and *Use model*. It is worth noting that, during the probability model's construction and application, these practitioners applied both mathematical thinking and statistical thinking. Mathematically speaking, they built the stochastic model based on mathematical assumptions and conditions. Statistically speaking, they constantly retested these mathematical assumptions and conditions against statistical data and how they were sampled to revalidate and rectify the existing model.

The interpretations of the confidence interval, especially the understanding of the confidence level, demand the knowledge of probability and have been demonstrated to be the most prominent misconception among both statistics students and professional researchers (Fidler & Loftus, 2009). For instance, Hoekstra, Morey, Rouder, and Wagenmakers (2014) listed seven statements based on the statistical result that "the 95% confidence interval for the mean ranges from 0.1 to 0.4" (p. 1163):

1. The probability that the true mean is greater than 0 is at least 95%.
2. The probability that the true mean equals 0 is smaller than 5%.
3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect.
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.

6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.
7. If we were to repeat the experiment over and over, then 95% of the time the confidence intervals contain the true mean. (p. 1160)

The first four statements are false because they have assigned probabilities to the parameter or the null hypothesis, which violates the basic constructs in a frequentist model. Statements 5 and 6 are false because they incorrectly provide a statement describing the property of the population mean solely based on a specific calculated interval, when, in fact, confidence intervals only "provide for a statement about the performance of the procedure of drawing such intervals in repeated use" (p. 1159). The last statement above is the correct way to interpret the confidence interval.

The notion of $p$-values is another vital aspect in introductory statistics where probabilistic thinking takes place. In a typical introductory statistics class, $p$-values are commonly taught as a numerical value used to identify statistically significant results. The mathematical rule of $p$-values states that if the $p$-value is less than the significance level, the null hypothesis is rejected, and the observed effect is statistically significant. Otherwise, the null hypothesis is not rejected, and the observed effect is statistically insignificant.

However, similar to confidence intervals, $p$-values are often misused or misinterpreted by students or even some statistics practitioners (Fidler & Loftus, 2009; Franklin et al., 2015; Motulsky, 2014; Reaburn, 2014; Wasserstein & Lazar, 2016). In 2016, to address the rising concerns of the validity and reproducibility of scientific studies in academia (Peng, 2015), the American Statistical Association published a statement on statistical significance and $p$-values (Wasserstein & Lazar, 2016). Particularly, they provided an informal definition for $p$-values:

> Informally, a $p$-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value. (p. 131)

as well as six principles to address common misconceptions about $p$-values:

1. $p$-values can indicate how incompatible the data are with a specific statistical model. (p. 131)
2. $p$-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. (p. 131)
3. Scientific conclusions and business or policy decisions should not be based only on whether a $p$-value passes a specific threshold. (p. 131)
4. Proper inference requires full reporting and transparency. (p. 131)
5. A $p$-value, or statistical significance, does not measure the size of an effect or the importance of a result. (p. 132)
6. By itself, a $p$-value does not provide a good measure of evidence regarding a model or hypothesis. (p. 132)

From the perspective of statistical thinking, $p$-values are more than plainly numbers for hypothesis rejection. A $p$-value tells us how surprising the observed result is when the desired effect is absent in the population (Franklin et al., 2015; Franklin et al., 2007). However, it does not tell us how strong the effect or the difference is since the $p$-value does not measure the magnitude of the studied effect (Hand, 1998; Motulsky, 2014). It is dangerous to make a decision merely based on the mathematical value of a $p$-value (Wasserstein & Lazar, 2016). A study that yields a $p$-value above the threshold, such as 5%, does not necessarily imply that there is no or little effect. Similarly, a low $p$-value does not ensure real statistical significance or even practical significance. According to the statistician and geneticist R. A. Fisher, who created the concept of $p$-values, a low $p$-value, such as the one less than 5%, simply suggests that the experiment should be repeated. If the new $p$-value falls below the threshold again, then the observed result is probably due to the desired effect rather than chance alone (Fisher, 1946).

Interestingly, Cobb and Moore (1997) analogized the classic chicken or the egg causality dilemma to the role of probability in introductory statistics:

> In the ideal Platonic world of mathematics, we can start with a probabilistic chicken and use deductive logic to lay a statistical egg, but in the messier world of empirical science, we must start with the egg as observed data and construct a prior probabilistic chicken as an inference. In an introductory statistics course, the chicken's only value is to explain where eggs come from. It seems a bit unfair, in that context, at least, to ask beginning students to learn about egg-generators before they've become familiar with eggs – less extreme, but in the same spirit as starting the study of chemistry with quantum mechanics. (Cobb & Moore, 1997, p. 820)

There is an on-going discussion on what topics in probability should be taught in an introductory statistics course (Franklin et al., 2015; Franklin et al., 2007; Moore, 1988). According to the GAISE report (Franklin et al., 2007),

> The concepts of probability needed for introductory statistics (with emphasis on data analysis) include relative frequency interpretations of data, probability distributions as models of population of measures, an introduction to the normal distribution as a model for sampling distributions, and the basic ideas of expected value and random variation. Counting rules, most specialized distributions and the development of theorems on the mathematics of probability should be left to areas of discrete mathematics and/or calculus.
>
> 1. Probability is an attempt to quantify uncertainty.
> 2. Particularly important to statistical inference is the notion of independence in the sampling settings.
> 3. the concepts of probability play a critical role in developing statistical methods that make it possible to make inferences based on sample data and to assess our confidence in such conclusions. (pp. 84-85)

Cobb and Moore (1997) recommend that, rather than teaching rules and theories of formal probability, introductory statistics teachers should spend valuable instructional time assisting students with understanding complex statistical ideas such as the sampling distributions. The instruction of probability should be fundamental and informal, with an emphasis on modeling and simulation. Fortunately, as technology becomes more advanced and accessible, more and

more instructors are teaching statistical inference with technological simulation and modeling (Cobb, 2007; Garfield, Zieffler, et al., 2012).

## Research on Statistical Thinking

In the following subsections, studies on statistical thinking and their significant findings are presented and organized according to the investigative steps in the GAISE framework (Franklin et al., 2007).

### Formulate Questions

Formulating or posing statistical questions, the beginning of a scientific, statistical investigation, has been long ignored by the school curriculum and receives little attention among statistics educators and researchers (Franklin et al., 2015). At almost all levels of beginning introductory statistics education, students are taught with ways to collect, summarize, organize, and describe data, and teachers are guided by the textbook to start the statistical discussion by posing questions and setting the context for students. It is not until recent years when the concept of statistical questions is accentuated in early chapters of school textbooks (Allmond & Makar, 2010).

Franklin et al. (2007) explained that it was challenging for novice statistics students to understand and appreciate "the difference between a question that anticipates a deterministic answer and a question that anticipates an answer based on data that vary" (p. 11) without introducing later stages, such as collecting and analyzing data, first. Arnold (2007) characterized these two types of questions as *survey* questions, "the question being asked to get the data" (p. 7), and *investigate* questions, "the question being asked of the data" (p. 7), respectively. In the subsequent discussion of this dissertation, the term "*statistical* questions"

will be used to equate "investigate questions". Based on Konold and Higgins (2003)'s study, novice statistics students usually viewed the data individually. The most common way to trigger them to view the data differently was by posing a question that asked students to compare group characteristics. It was then perhaps evident to students that data could also be viewed aggregately and that their view of data shall depend on the type of question raised. Another way, more straightforwardly, was to provide examples and guidance in two different types of questions that students could raise. Through this way, students were found to pose more statistical questions, including those that compared group characteristics, that could be addressed with a statistical investigation (Allmond & Makar, 2010; Chin & Kayalvizhi, 2002). It was also detected that both students (Arnold, 2007) and many statistics teachers (Bargagliotti et al., 2014) lacked judgment in what contributes to a good statistical question. Their ability to formulate questions mainly stayed at the level of replicating what they had been exposed to solely (Lavigne & Lajoie, 2007).

**Collect Data**

The concept of samples and sampling involves the notions of population, sample, sample size, sampling method, sampling distribution, and, most importantly, sampling variability (Pfannkuch, Arnold, & Wild, 2015). When choosing a sound sampling method, most students applied their intuition on whether or not the selection criteria sounded fair to all population members. Very few of them focused on the actual probability of an individual being selected (Jacobs, 1999; Meletiou-Mavrotheris & Paparistodemou, 2015). Such a finding could be discerned as a direct result of how many introductory statistics concepts are usually presented and delivered in the school curriculum. Contrary to mathematics, current introductory

statistics purposely deludes the theoretical part and promotes the practical purpose of statistical concepts, as recommended by many statistics researchers and educators (Cobb & Moore, 1997; Hand, 1998; Moore, 1988). In college, introductory statistics is a required course for many undergraduate majors, many of which are non-STEM. It makes sense that textbooks are typically written for a broader range of audiences with less emphasis on mathematics and more attention to its practical implication (Carver et al., 2016).

"In statistics the purpose of a sample is to show the variation in a population so it can be characterized and summarized" (Watson, 2013, p. 28). Yet, students, and possibly some teachers as well, do not perceive this way. Saldanha and Thompson (2002) conducted a teaching experiment on 27 high school AP statistics students and differentiated two conceptions of how students understand the concept of a sample:

1. Additive: a sample is a subset of the population. Multiple samples are multiple subsets of the population. Characteristics of a sample are solely based on the sample selected and independent of characteristics of the population.
2. Multiplicative: a sample is a "quasi-proportional mini version of the sampled population" (p. 266). Multiple samples are "multiple, scaled quasi-mini-versions of the population" (p. 267). The characteristics of a sample are expected to vary but remain within bounds around the population's corresponding characteristics.

When covering statistical inference in an introductory statistics course, many instructions students have received or are exposed to are based on the information obtained from a single sample. They could appear in textbook examples, homework, project, and even exams. Such a learning model limits students' understanding of a sample to additive only and hinders their development of the conception of sampling variability. Lee and Hollebrands

(2011) claimed that modern technology, such as Tinkerplots, could easily simulate the statistical process of resampling and effectively promote students' aggregate-based reasoning. But based on investigations in Tinkerplots conducted by others, despite the dynamic data representation and sampling animation, young students were unaware of the underlying connection between sampling variability and population variation (Gil & Ben-Zvi, 2010; Saldanha & McAllister, 2014).

In contrast, college students are forcefully instructed to "accept" such a connection, mostly by formal definitions. An example of such a definition is the most important theory in introductory college statistics: the central limit theorem (CLT). Unfortunately, students exhibit misconceptions regardless. According to Yu et al. (1995),

1. Students often perceived random sampling as a self-correcting process in which the final sample will, for sure, include all characteristics of the population. However, they failed to recognize the complexity in real-life random sampling, which is all about the probability independence among samples being collected.
2. Students tended to believe that a sample distribution or a sampling distribution should always conform to the population distribution shape. They also confused a sample distribution with a sampling distribution.
3. Students often drew the bell-shaped normal curve as an inverse letter "U" without realizing the asymptotic tails at the two ends that extend indefinitely and never intercept with the horizontal axis. The use of graphical display on a computer failed to clear up such a misconception due to technology restraints.

4. Students cared too much about the minimum sample size requirement for the CLT. They failed to detect that such a cutoff value is never fixed and could be easily influenced by the degree of non-normality the population could have.
5. Students were unable to comprehend hypothesis testing in terms of the CLT. They failed to acknowledge that hypothesis testing is a probabilistic inference based on the theoretically infinite samples taken from the population with replacement. As a result, they believed that the single sample selected reflects all characteristics of the population.

Researchers and educators advocated that, when introduced with the concept of samples and sampling, regardless of the level, students should be provided with meaningful learning experiences that mimic the data collection in a genuine statistical analysis (Meletiou-Mavrotheris & Paparistodemou, 2015; Moore, 1988). When available, with careful guidance from teachers, dynamic statistical software should be implemented to visualize and clarify some of the more abstract and intricate aspects of samples and sampling (Biehler, Ben-Zvi, Bakker, & Makar, 2012; Konold, 2007; Watson & Donne, 2009).

**Analyze Data**

School data analysis usually begins with transnumeration, typically visual representations of the data, that engenders a better understanding of what is being dealt with (Wild & Pfannkuch, 1999). Students begin with the formalization of various types of graphical methods, either by hand (Chick et al., 2005) or with physical manipulatives (Chick & Watson, 2001; Madden, 2008). As the dataset expands, statistical graphing technologies are introduced to reduce manual transnumeration of large datasets. Particularly, dynamic technological tools, such as Tinkerplots and Fathom, have shown promising results in improving students'

conceptual understanding of topics in statistics and probability (Biehler et al., 2012). Pea (1987) recapitulated that technology is used to "amplify" learner's problem-solving ability and "reorganise" their ways of thinking during problem-solving. However, it is not always the case that using these tools will automatically improve user's conceptual understanding (Fitzallen, 2013; Zbiek et al., 2007). For more discussion on this matter, see section Statistical Teaching with Technology on page 43.

Studies on visual representations have been focusing on students' diagrammatic reasoning. For instance, according to Capraro, Kulm, and Capraro (2005), students found the need to graph zero above the $x$-axis. "Students explained that leaving a number without some representation did not seem correct" (p. 167). Students also had a hard time telling the difference between a bar graph and a histogram. Additionally, students were not able to use the graph they constructed to interpret what a typical value means. Bakker et al. (2006) investigated students' graphical interpretation of centers by connecting to variation between and within data sets in Tinkerplots and found out that novice statistics students usually used informal statistics language, such as "clump" and "longer shape," along with rudimentary diagrammatic reasoning, to describe and interpret graphs. Additionally, students often relied on these informal reasonings to compare and check their formal computations of descriptive statistics. The study concluded that introductory statistics instruction should begin with exploratory and informal analysis in order to develop initial notions of center and variation. By rationalizing what they see in words and representing what they think in graphs, students can learn to perceive statistics like statisticians do. Recent work by Konold and Harradine (2014), Makar (2014) and Watson, Chick, and Callingham (2014) in the related area highlighted the connections between descriptive statistics and informal inference via repeated sampling as well

as how these connections can enhance students' overall learning experience in and conceptual understanding of introductory statistics.

However, there is a lack of current teachers' knowledge about statistical graphs (González, Espinel, & Ainley, 2011). Most studies focused on international prospective teachers only. It was found that these teacher candidates experienced the same difficulty as statistics students when recognizing (González & Pinto, 2008), constructing (Bruno & Espinel, 2009), comparing (Burgess, 2002), and interpreting (Batanero, Arteaga, & Ruiz, 2010; Espinel, Bruno, & Plasencia, 2008; Rouan, 2002) common statistical graphs. Among a few studies which demonstrated positive findings (Monteiro & Ainley, 2006, 2007), the statistical knowledge tested, though, was relatively low.

In addition to graphical representations of the data, the data need to be simplified and summarized with descriptive statistics before unveiling more meaningful information. Like the two different views on samples in terms of the population, descriptive statistics switch the view of the data from individual to aggregate through measures of center, dispersion, and relative standing.

There are abundant early studies on students' conceptual understanding of central tendency and representativeness. Mokros and Russell (1995) interviewed 21 fourth-, sixth-, and ninth-graders and summarized five approaches used by young students for understanding the term "average:"

1. Use mode but ignores the rest of the data values.
2. Use mean but only refer to its mathematical procedure.
3. Use mean or a self-reported representative value of the data with scrutinization for reasonableness.

4. Use median or midrange.
5. Use mean and perceive it as a balancing point of a data set.

Students using the first two approaches viewed average as a mathematically produced number that had no meaningful purpose for describing the data set. On the contrary, students using the last three approaches realized what it meant to select a representative value of a data set and understood why the chosen value made sense to them for describing the center of a data set. During a comparative study, Cai (2000) examined the knowledge of arithmetic mean from 232 American sixth-graders and 311 Chinese sixth-graders. When reasoning their problem-solving process, Chinese students preferred using abstract mathematical thinking and solving algebraically, whereas American students made connections to real-life objects and explained either visually or verbally. Even though Chinese students outperformed their American counterparts in finding the final correct answers, both groups of students failed to view arithmetic mean as a meaningful characteristic describing the center of a data set, which limited all participants' problem-solving strategy to mainly the mathematical formula for arithmetic mean. Pollatsek, Lima, and Well (1981) interviewed 17 undergraduate students, most of whom were psychology majors. All participants were asked to "think out loud" while working on weighted mean problems. It was found out that these students had difficulty calculating the weighted mean, especially when weighting and combining multiple means into a single mean. Similar to the findings for younger students, these college students lacked the notion of representativeness. Unfortunately, not much of students' conception of central tendency has been improved since those findings from early investigations (Makar, 2014).

Surprisingly, statistics teachers' understanding of descriptive statistics is no better than their students'. For instance, in a study conducted by Jacobbe and Horton (2010), elementary

teachers with a strong mathematical background were shown to lack higher levels of data display comprehension, i.e., reasoning center and variability with multiple box plots. Groth and Bergner (2006) investigated 46 pre-service elementary teachers' understanding of central tendency and discovered that, though these teachers were able to differentiate and compute mean, median mathematically, and mode correctly, they had difficulty explaining what these measures of center represented from a statistical point of view. Similar findings were replicated by Jacobbe (2012). At the secondary level, Hammerman and Rubin (2004) examined strategies that teachers applied to describe data variability when comparing groups in Tinkerplots and reported that teachers "tended to view slices of the distributions out of context of the rest of the distribution" (p. 37) and were unable to compare variability among different distributions. Similarly, Makar and Confrey (2004) asked teachers to compare the performance of different types of students in Fathom and reached the same conclusion as Hammerman and Rubin (2004)'s. Additionally, they reckoned that most teachers also lacked appropriate statistical language when describing and comparing distributions, which was later concurred by Hannigan et al. (2013). However, through reflecting on students' misconceptions, teachers' understanding of descriptive statistics tended to grow as they gained more teaching experience (Cai & Gorowara, 2002).

Another popular area of research in data analysis is correlation and regression. Sorto, White, and Lesser (2011) investigated students' conception of the line of best fit and concluded that the least-squares method did not occur to students naturally. Below are some reported conceptions from students regarding the placement of the line of best fit:

1. The line passes as many points as possible.
2. The line passes the origin and separates an equal number of points above and below.

3. The line connects the left-most point and the right-most point in the scatterplot.
4. The line goes through the main cluster of points.

A study conducted by Casey and Wasserman (2015) revealed similar criteria for selecting the line of best fit from mathematics teachers. Both Moritz (2004) and Bakker (2004) reported that students tended to lack an aggregate view of the data points, which prevented them from understanding the least-square method as well as identifying the overall pattern and relationship between bivariate data. To resolve it, Dierdorp, Bakker, Eijkelhof, and van Maanen (2011) and Shaughnessy (2007) suggested that teachers should assign open-ended exploratory tasks that required students to create a linear model for their own collected data. Such activities would encourage students to seek and reflect on their solutions, promoting deeper statistical thinking.

**Interpret Results**

The concept of statistical inferences has been a challenging topic for students at all levels (Garfield & Ben-Zvi, 2008). Partially, it is due to the mathematical nature of various procedures that students need to identify, follow and apply (Moore, 1988). Meanwhile, unlike other topics in statistics, such as descriptive statistics, which have been exposed to students multiple times throughout all grade levels, the concept of statistical inferences is relatively new to many students when it is first officially introduced (Pfannkuch, 2005). Additionally, students' practice on statistical inferences often remains within the scope of textbook practice problems isolated from other steps in the statistical investigative cycle. Students complete these exercises by treating them as mathematical problems: remove distracting contexts, symbolize given information, select the appropriate model (which is usually given), plug into

formulas, and compute (Bakker & Derry, 2011; Shaughnessy, 2007). From many students' perspectives, statistical inferences, such as confidence intervals and significance tests, are merely mathematical recipes that are static and stale (Cobb & Moore, 1997; Makar & Ben-Zvi, 2011). Lastly, and probably more undeviatingly, teachers being inexperienced in statistical inferences could be the primary reason for students' low retention in statistical inference. Both Bargagliotti et al. (2014) and Hannigan et al. (2013) stated that prospective teachers had false beliefs in statistical inferences and inferior understanding of sampling variability along with its relation to inferential statements.

Many studies have been conducted to investigate students' and researchers' interpretations of inferential statistics at the college level, such as the confidence interval and $p$-value. They revealed many misconceptions. Kalinowski et al. (2010) surveyed 94 students in science disciplines and discovered that about 25% of the respondents falsely believed that as the confidence level increased, the width of the confidence interval would decrease. Additionally, about a third of the respondents falsely believed that the underlying distribution of a confidence interval was uniform rather than bell-shaped. In a study conducted by Hoekstra et al. (2014), prevalent misconceptions about the interpretations of confidence intervals were identified among 596 psychology students and faculty, regardless of their previous training experience in statistics. However, not all studies revealed worrying results. Some studies attempted to seek approaches to rectify some of the common misconceptions about inferential statistics. For instance, another widespread misconception about confidence intervals was the equivalence between statistically insignificance and "no effects" (Haller & Krauss, 2002; Oaks, 1986). Fidler and Loftus (2009) conducted two experiments on statistics students and concluded

that "[v]isually represented confidence-intervals substantially alleviated the misinterpretation of statistically nonsignificant results as evidence for the null hypothesis" (p. 36).

In an attempt to improve the quality of instruction in teaching statistical inference, some studies concentrated on the advocation of informal inferential reasoning (Gil & Ben-Zvi, 2011) combined with exploratory data analysis (Cobb, McClain, & Gravemeijer, 2003) prior to introducing students with any formal statistical inference. To better understand how students make decisions when performing data analysis, some studies characterized different types of reasonings from students during the complete statistical investigative process (Lavigne & Lajoie, 2007; Watson & English, 2015). However, despite Fielding-Wells (2010) having introduced explicitly the PPDAC statistical investigative cycle (Wild & Pfannkuch, 1999), students still encountered substantial difficulty linking different steps in a statistical analysis. A few studies recommended that explicit instructions be offered to backtrack how certain conclusions are reached by connecting them to problems, data, and evidence. Exposing students to later stages in the statistical investigative cycle may enable students to see the goal of the statistical analysis and the achievement of concluding on a question that anticipates variability (Fielding-Wells & Makar, 2015; Fitzallen, Watson, & English, 2015). Meanwhile, teacher education programs should provide teacher candidates with more hands-on experience in a complete statistical investigation (Madden, 2011; Santos & da Ponte, 2014).

In 2015, the NCTM published a book outlining five big ideas in introductory statistics education, which are used to summarize this section on statistical thinking:

> Big Idea 1. Data consist of structure and variability.
> Big Idea 2. Distributions describe variability.
> Big Idea 3. Hypothesis tests answer the question, "Do I think that this could have happened by chance?"

Big Idea 4. The way in which data are collected matters.
Big Idea 5. Evaluating an estimator involves considering bias, precision, and the sampling method. (Crites & Laurent, 2015, pp. 127-128)

## Statistical Teaching with Technology

### Statistical Technologies

There have been many publications on various types of educational technology available for statistics education in the past two decades. Pea (1987) defined *cognitive technologies* or *cognitive tools* as those technologies that potentially "transcend the limitations of the mind . . . in thinking, learning and problem-solving activities" (p. 91). To include a wider variety of technology, Biehler (1997) classified technology related to the teaching and learning of statistics into *tools* (i.e., Excel with statistical packages, graphing calculators), *resources* (i.e., downloadable datasets), and *microworlds* (i.e., interactive environments such as Tinkerplots and Fathom) in order to distinguish technologies that merely perform statistical analysis and provide datasets from those that support interactions between the technology and the user. Similarly, Bakker (2002) specified Biehler's tools as *route-type tools* and microworlds as *landscape-type tools.* Alternatively, Zbiek et al. (2007) differentiated technology based on the type of mathematical activity: technologies that help conduct *technical mathematical activities* that involve procedural and mechanical calculations and technologies that help perform *conceptual mathematical activities* that require examination, validation, and reflection.

Nevertheless, from the diverse classifications of technology, we infer that technology implementation is far beyond a binary decision–to implement or not to implement. There are sophisticated factors to be taken into account before any implementation. According to Ertmer

(2005), two types of factors could impact whether and how teachers implement technology in class:

- external: professional training for using technology, technology availability in a class, and school support;
- internal: teacher's knowledge about technology, beliefs in technology, and preparation for teaching using technology.

Ruggiero and Mong (2015) added a third factor that mixes both external and internal factors: the degree to which the technology integration matches teacher's pedagogical goal and teaching style. The addition of integrating technology into statistics education definitely brings more challenges to mathematics teachers.

The first possible concern that comes to mind to most mathematics teachers could be *feasibility*, such as: Do teachers know about current technology options for teaching statistics? Is the technology affordable for schools or students? How many platforms (i.e., PC, MAC, Android, iOS) or languages does the technology support? Does the technology require a constant internet connection? What is the learning curve of the technology for teachers and students? Is the user interface intuitive to teachers and students? With a chosen technology, what and how can statistical topics be taught? How long will it take for an experienced user to design a statistical task? Are there any community support systems for professional development?

In addition to instructors' concerns about the convenience of implementing technology in teaching statistics, some researchers and mathematics educators dispute that technology may not accomplish its pedagogical purposes for various reasons. For instance, Yu et al. (1995) pointed out that many computer programs and web applets failed to illustrate the property of a

normal distribution, which states that the normal curve extends indefinitely to both ends but never touches the horizontal axis. The limited and discrete display of the distribution on computers also constrained students' interpretation and understanding of the concept of infinity. Dick (2008) described this phenomenon using the term *mathematical fidelity*, which indicates that technology sometimes incorrectly represents the underlying mathematical object. Often, it is due to the limitations of technology or the design choice for a friendlier user-interface.

Even though the appropriate use of technology tremendously fosters the development of sense-making in mathematics (Chance, Ben-Zvi, & Garfield, 2007; Chance & Rossman, 2006; Cobb, 1994; Franklin et al., 2015; Franklin et al., 2007; Shaughnessy, 2007), it does not automatically grant mathematically meaningful feedback to the user–a connection or relationship must be built between students and the technology they use. Forming a relationship between an instrument, either physical or virtual, and the user is considered as the characterization of *instrumental genesis*, which, in mathematics education, relates to the situation where the student realizes the mathematics underlying the instrument (*instrumentation*) and uses it for his or her own mathematical purpose (*instrumentalization*) (Verillon & Rabardel, 1995). Regardless of the type of technology used, if *instrumentation failure* occurs, which refers to the situation where a student treats the technology task as some tangible replacement of the traditional pedagogy but fails to recognize the underlying abstract mathematical object, then *instrumentalization failure* transpires, meaning there will not be any meaningful learning outcomes produced by implementing and using the technology. This results in a low degree of *cognitive fidelity* when student's conceptual understanding or method

of solution is barely resembled by the technology's explicit representation of the pertaining mathematical object or its form of solution (Zbiek et al., 2007).

Pedagogically speaking, to optimize the statistical learning experience with technology, an experienced mathematics teacher needs to know more than statistical content knowledge to guide and assist students through extracting statistically meaningful feedback from interactions with the technology. In general, depending on the goals of teachers' as well as students', there are two types of mathematical activity a teacher can conduct to enhance students' learning experiences–*exploratory activity* and *expressive activity* (Zbiek et al., 2007). These two activities could be applied to technology-based teaching and learning as well. During a technology-based exploratory activity, teachers often give students a procedure to follow with a clear goal. The procedure could be specific or general. For example, both of the following two instructions aim to find the right cumulative area underneath a normal curve using the same web applet, but the instructions are given differently.

- In this web applet, enter the mean and the standard deviation of all heights of your gender as well as your height in the text boxes. Hit this button to calculate the cumulative area to the left of your height. Then use the fact that the total area under the normal curve is 1 to find the cumulative area to the right of your height. Convert it to a percentage and report it as the percentage of people of your gender taller than you.
- Using this web applet to determine the percentage of people of your gender who are taller than you.

During an expressive activity, however, teachers often provide no instructions on specific tasks. Students are free to explore the concept by taking advantage of all the technology has to offer. For instance, the teacher could ask students to create a technology-based task on the

central limit theorem for someone who has trouble understanding that concept. Sometimes a mathematical activity could be both exploratory and expressive with equal or unequal emphasis on either part. Despite the type of mathematical activity, the teacher could act as a technical counselor whose goal is to resolve any technology-related issues and improve students' statistical literacy of using modern technology. Moreover, the teacher could also fulfill her duty as a statistics instructor whose goal is to assist students with any statistics-related questions and enhance their conceptual understanding.

Nonetheless, in reality, technology-based statistical activities do not always go as planned. There could be mismatches between the technological task and the teacher's pedagogical goal. For instance, when applying technology, students may view using technology as a replacement for memorizing and understanding statistical formulas and become gradually dependent on and blindly-faithful to technology-produced results. Dick (2008) and Zbiek et al. (2007) characterize such mismatch by the term *pedagogical fidelity* which denotes the extent to which the technology fulfills the pedagogical goals, the extent to which the teacher believes in the technology in furthering students' learning, and the extent to which the students recognize the technology as something beneficial to their study.

In an endeavor to resolve a potentially low degree of various fidelity issues in statistics education, Madden (2008) proposed a pedagogical model called the *Dynamic Technology Scaffolding* to support teachers' development of instructional tasks for teaching statistics. The scaffolding starts with a physical exploration of the concept using concrete materials or physical manipulatives so that students know the basis behind a new concept. It is followed by a virtual exploration of the same concept in a pre-designed and simulated technological environment to help students make connections and conjectures. The scaffolding ends with

students reconstructing and manipulating previous physical or virtual explorations in the hope of gaining a higher level of conceptual understanding.

**Recommended Knowledge**

The term "technology" may refer to many things, depending on the field of discussion and the time of speaking. In this study, technology mainly refers to applicable digital technologies, applications that run on computers, tablets, and mobile devices. According to Koehler and Mishra (2009), technologies commonly share three properties:

1. *protean* – they are able to accomplish many different things, which creates a learning curve for many users;
2. *unstable* – they are frequently updated and replaced with newer and better versions, which means that sometimes the user must start over with another different learning curve;
3. *opaque* – their fundamental affordance and constraints cannot be modified by a typical user, limiting what a user can do and making it harder to select a suitable option for the user's needs.

Due to these common properties of technologies, the implementation of educational technology in classrooms is particularly challenging to many teachers. Teachers need to learn how to use technology as well as how to teach with technology. This subsection includes several theoretical frameworks on teacher's knowledge of teaching using technology. Particularly, they were selected for their close connection to introductory statistics.

The technology, pedagogy, and content knowledge framework, or the TPACK framework (Figure 1.1), was built based on the construct of the pedagogical content knowledge

(PCK) by Shulman (1986, 1987), with the additional inclusion of technological knowledge. In the study of this dissertation, there are three levels of knowledge components in TPACK: *individual*, *between*, and *among*. At the "individual" level, we have

- *Technological Knowledge* (TK): the knowledge of technologies in statistics;
- *Pedagogical Knowledge* (PK): the knowledge of teaching and learning in general;
- *Content Knowledge* (CK): the knowledge of statistics necessary for teaching.

At the "between" level, we have

- *Technological Pedagogical Knowledge* (TPK): the knowledge of using available technologies to teach statistics effectively;
- *Technological Content Knowledge* (TCK): the knowledge of selecting suitable technologies for teaching statistics;
- *Pedagogical Content Knowledge* (PCK): the knowledge of teaching and learning in statistics.

At the "among" level, we have *Technological Pedagogical Content Knowledge* (TPACK): the knowledge "that emerges from interactions among content, pedagogy, and technology knowledge" (Koehler & Mishra, 2009, p. 66). TPACK tests teacher's ability in five steps:

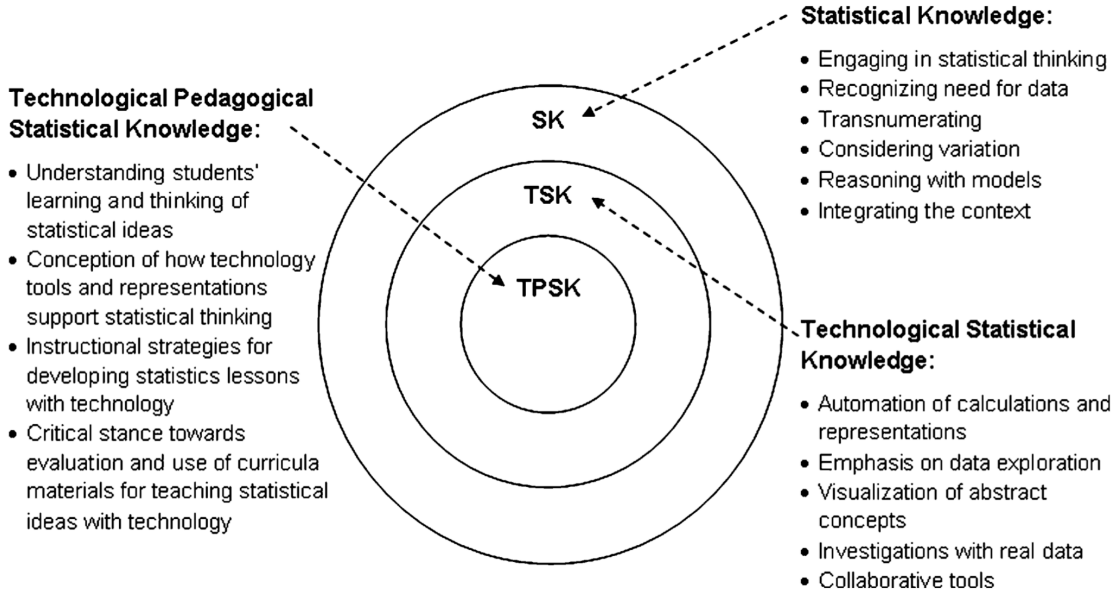
1. recognize the underlying statistical concept and represent it correctly using technology;
2. present technological representation of the underlying statistical concept in a soundingly pedagogical way that is easier for students to understand;
3. address students' common difficulties or misconceptions via technological representation;
4. connect technological representation with students' prior learning experience;
5. develop students' new epistemologies and reinforce their old ones.

Different from classifying knowledge into seven different domains, Lee and Hollebrands (2011) identified three "layered" domains (Figure 2.2):

- *Statistical Knowledge* (SK): the knowledge of statistics with emphasis on statistical thinking.
- *Technological Statistical Knowledge* (TSK): the knowledge of recognizing technologies for statistical teaching and conducting real-life statistical analysis using technology.
- *Technological Pedagogical Statistical Knowledge* (TPSK): the knowledge of selecting and using appropriate technologies to represent statistical concepts and promote statistical learning.

Visually, SK is the outermost and largest domain enclosing TSK and TPSK, indicating that statistical knowledge is fundamental for teachers to learn TSK and TPSK. TSK is layered in the middle, serving as another necessary but not sufficient predecessor for developing TPSK.

Figure 2.2. *The TPSK Framework and Its Layered Three Domains. Reprinted from Lee and Hollebrands, 2011, p. 362.*



Groth (2007) characterized the "common knowledge" as well as the "specialized knowledge" required for teaching statistics based on the four investigative steps in statistical problem-solving by Franklin et al. (2007). Common knowledge refers to the knowledge one develops in traditional statistics courses, which may require specific mathematical competency. Specialized knowledge refers to the knowledge one needs in order to teach statistics effectively. To distinguish the knowledge used for teaching statistics, examples of tasks were given for those requiring primarily statistical knowledge and mathematical knowledge, respectively (see Table 2.1). The author concluded that mathematics is a distinct discipline from statistics but is applied and used in many aspects of statistics.

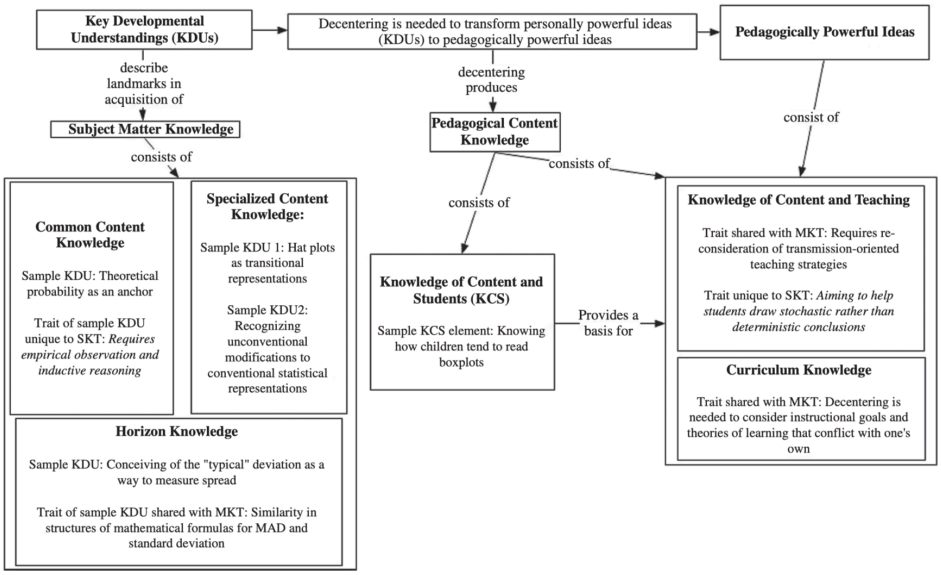
To generalize those hypothesized aspects of statistical knowledge for teaching, Groth (2013) proposed a hypothetical framework that characterizes the Statistical Knowledge for Teaching (SKT) into two board domains: Subject Matter Knowledge and Pedagogical Content Knowledge (see Figure 2.3). The framework starts with identifying Key Developmental Understandings (KDUs) in statistical learning – "cognitive landmarks in learning subject matter" (p. 127) that significantly advance students' conceptual understanding in statistics. These KDUs describe the necessary statistical knowledge needed for teaching, which is categorized as the first domain of SKT – the subject matter knowledge. Within the subject matter knowledge, there are three subdomains:

- *Common Content Knowledge*: "knowledge that teachers are responsible for teaching their students" (p. 123).
- *Specialized Content Knowledge*: "knowledge of various representations and unusual student strategies" (p. 123).

Table 2.1. *Hypothesized Aspects of Statistical Knowledge for Teaching I. Reprinted from Groth, 2007, p. 430.*

| GAISE framework component | Knowledge type | Examples of tasks requiring primarily mathematical knowledge | Examples of tasks requiring primarily nonmathematical knowledge. |
|---|---|---|---|
| Formulating questions | Common | Accurately reading a box plot in order to formulate questions from data | Understanding the difference between a deterministic and a stochastic question |
| | Specialized | Understanding differences between how students read box plots and dot plots | Appraising the potential fruitfulness of student-posed questions |
| Collecting data | Common | Constructing accurate simulation algorithms and measuring quantities correctly | Constructing survey questions and designing experiments |
| | Specialized | Understanding students' strategies for measurement and identifying difficulties students might have constructing simulation algorithms | Anticipating students' difficulties in distinguishing the roles and purposes of random sampling from those of random assignment |
| Analyzing data | Common | Computation of descriptive statistics such as mean, median, and mode | Navigating "typical value" and "signal-in-noise" statistical contexts |
| | Specialized | Identifying the mathematical properties of the mean that can be difficult for students to comprehend | Realizing that students may compute the arithmetic mean for a data set without regard for the context of the data |
| Interpreting results | Common | Correctly interpreting the mathematical meaning of the concept of $p$ value | Judging the appropriateness of a significance level chosen by a researcher |
| | Specialized | Understanding students' interpretation of the concept of $p$ value | Anticipating students' over-generalization of the term *significant* |

Figure 2.3. *Hypothesized Aspects of Statistical Knowledge for Teaching II. Reprinted from Groth, 2013, p. 143.*




- *Horizon Knowledge*: "knowledge of the broader content landscape in which curriculum is situated" (p. 123).

Through "decentering", a process in which teachers switch their perspective on statistics from themselves to students, teachers can transform their KDUs into "pedagogically powerful ideas" that consist of the second domain of SKT – the pedagogical content knowledge. Within the pedagogical content knowledge, there are three subdomains:

- *Knowledge of Content and Students*: "knowledge of common student difficulties with content and their thinking patterns" (p. 123).
- *Knowledge of Content and Teaching*: "knowledge of content-specific teaching strategies" (p. 123).
- *Curriculum Knowledge*: "knowledge of how to arrange curricula to enhance student learning" (p. 123).

# Chapter 3

# METHODOLOGY

## Participants

Fifteen mathematics teachers from a local community college in New York City were selected to participate in this study. They were chosen mainly because of three reasons.

First, they were selected from a group of mathematics instructors who were more than likely to have taught introductory statistics more than once due to the high demand for introductory statistics sections. In Spring 2020, the college offered 89 sections of introductory statistics courses and 46 co-requisite sections of introductory statistics with the integration of college algebra. For Fall 2020, there were 100 and 40 such sections available to students, respectively. The capacity of these sections ranged from 25 to 35. For an average enrollment of 20 students per section, there were approximately 2700 introductory statistics students in the spring of 2020 and 2800 prospective introductory statistics students in the fall of 2020 at this institute alone. To cope with the vast demand of statistics instructors at the college, the mathematics department consisted of 76 full-time faculty and 160 adjunct faculty as of Summer 2020. Despite a large number of faculty members, many of the instructors teaching these sections were teaching multiple sections per semester and had been teaching the same course for many semesters. As a matter of fact, among the 15 selected adjunct faculty, 14 of them claimed to have taught introductory statistics for more than one semester. The remaining participant was teaching introductory statistics for the first time but was assigned with two sections at the time of the study (amid Spring 2020) and had taught introductory statistics in

other colleges previously. Even though no full-time faculty were recruited for the study, adjunct faculty contributed to the major teaching force of introductory statistics courses at the college. In Spring 2020, 51.7% (46 out of 89) of all introductory statistics courses and 65.2% (30 out of 46) of all co-requisite sections of introductory statistics were taught by adjunct faculty. Thus responses based on participating adjunct faculty could still render meaningful data to the research questions. Meanwhile, selecting participants with rich teaching experience in introductory statistics facilitated identifying key developmental understandings and pedagogically powerful ideas.

Second, they were selected from a group of mathematics instructors whose academic degrees mainly focused on mathematics instead of statistics. According to the department faculty page, as of Summer 2020, among 76 full-time faculty, there were 20 full professors, seven associate professors, 18 assistant professors, 25 lecturers or instructors, one continuing education teacher, and five lab technicians. According to the 40 online faculty pages that displayed complete degree information, 15 were PhDs in mathematics or applied mathematics, 17 were PhDs or EdDs in mathematics education, and only one stated PhD in mathematics and statistics. There was no full-time faculty member who reported having received any degree in statistics or statistics education. Yet, 67.5% of them (27 out of 40) wrote that they had taught introductory statistics course. Unfortunately, there was not enough information to summarize the highest degrees of adjunct faculty members since most of their pages were incomplete or empty. However, based on this study's demographic survey, 17 out of 19 were adjuncts who had taught introductory statistics, only one received a PhD in mathematics and statistics education. Selecting participants with little statistical but strong mathematical background created a hotbed for a potential conflict between mathematical thinking and statistical thinking

among participants. It was intriguing to exam whether mathematics teachers teach statistics with a mindset of mathematical thinking, statistical thinking, or both.

Third, they were selected because they were not restricted to using specific statistical technologies for their teaching. On the department syllabus posted online, the only technology required was a scientific or graphing calculator. Instructors were free to use any other technologies or no technologies (except the calculator), allowing potential participants to have various technology implementations.

## Procedures

All participants went through two stages and four phases.

### Stage 1: Before the interview

A recruitment email was sent to all 236 mathematics department faculty starting November 2019. Till May 2020, a total of 24 instructors responded to and inquired about the study via either email or phone call. Five instructors decided not to proceed after learning more about the study. For those who did give a reason why they did not want to proceed, one reason was that the compensation was not attractive, and the other reason was that they had to deal with the COVID-19 pandemic and did not have time to participate.

#### *Phase I: Technology Evaluation*

A link to an online survey (Appendix A) was then sent to those 19 instructors who agreed to proceed. The survey collected demographic information and teaching experience from participants. It simulated a scenario in which teachers try to learn about potential statistical technology for their teaching in real life. By going through the survey, participants

were assumed to gain a general knowledge of five technology options available for teaching statistics. These options were selected based on a comprehensive categorization by Chance et al. (2007) and currently available teaching resources that teachers could find independently.

1. *Minitab* (Statistical Software Package) is a professional statistical software used by statisticians, business analysts, teachers, and researchers. Minitab recommends rich teaching resources on its official website.
2. *Tinkerplots* (Educational Software) is a well-documented software in mathematics education literature that offers a dynamic learning environment for introductory topics in probability and statistics. On its official website, there are exploratory activities along with their Tinkerplots files that teachers can download and explore right away.
3. *Google Sheets* (Spreadsheets) is a free online program that can be molded for teaching purposes. For instance, Ling (2018) has created an online introductory statistics course using Google Sheets.
4. *Rossman/Chance Applet Collection* (Applet/Stand-alone Applications) is a list of web applets designed for randomization-based curriculum on introductory statistics topics by Rossman and Chance (2014).
5. *Rice Virtual Lab in Statistics* (Multimedia Materials) is a free online platform that combines electronic textbook, animated activities, videos, applets for demonstration and exploration, embedded exercises and links to resources for college-level introductory statistics. The website even allows users to download all applets and their source code with no restriction.

Participants were instructed to explore the general information about each option as well as any teaching resources available for statistics instructors. It was up to each participant how

long they wanted to spend and how deep they wanted to investigate each technology option. At the end of the survey, participants were asked to rate each technology option and provide comments on the pros and cons they observed. This survey was conducted before the interview day to ensure that all participants had similar general knowledge about different types of technology options in statistics, such as what they are, what they can do for statistics, and what they can do for teaching statistics. It also served as an inspiration for experienced teachers who might lack the recognition of available technology options for statistical teaching. However, upon completing this survey, participants were not required to know how to use these options.

Two full-time faculty and 17 adjunct faculty participated in the first stage of the study, and 16 of them, who were all adjunct faculty, moved on to the second stage. Among those three who did not continue, two of them were interrupted due to the COVID-19, and the other one decided not to continue because of personal reasons.

*Phase II: Statistical TPACK survey*

A link to a second online survey (Appendix B) was sent to those 16 participants who decided to continue. The survey was adapted from an existing one made by the creators of the TPACK framework (Schmidt et al., 2009) and aimed to evaluate participants' TPACK knowledge in introductory statistics. The version used for this study, however, had been slightly modified. First, the wording was changed to emphasize that the content of teaching was introductory statistics. Second, in the instruction, the definition of technology was clearly defined to correspond to what participants had explored online during Phase I. In this study, technology was referring to

- statistical software packages such as SPSS, SAS, R, Minitab;
- educational software such as Tinkerplots and Fathom;
- spreadsheets application such as Excel and Google Sheets;
- applets (online or offline) such as the Rossman/Chance Applet Collection that participants have explored;
- multimedia materials such as Rice Virtual Lab in Statistics that participants have read about;
- computers, tablets, interactive whiteboards, or any other technology facilitates teaching in the classroom.

Participants rated each statement based on their attitude and belief where 5 points represented "strongly agree," 4 points symbolized "agree," 3 points stood for "neither agree nor disagree,""not applicable," or "unsure," 2 points corresponded to "disagree," and 1 point was for "strongly disagree." In the end, descriptive statistics of points were computed for each participant across all seven domains of the TPACK framework.

**Stage 2: On the interview day**

After stage 1 was complete, each participant was contacted again to schedule a one-on-one interview. Each interview lasted for approximately 90 minutes and was audio-taped. Due to the COVID-19 pandemic, only the first five interviews were conducted in the adjunct faculty office where all 16 participants worked. The 11 remaining interviews were administered online via the video conference application *Zoom*. However, one interviewee could not complete the interview due to technical difficulties. In the end, a total of 15 participants, who were all part-time statistics instructors, successfully completed the study.

During each in-person interview, every participant was provided with a tablet and a stylus. An interview handout was distributed electronically via a tablet at the beginning of the interview. A note-taking application called *Notability* was used to record the interview audio and participant's handwriting on the tablet. For online interviews, a screen of the handout was shared with the participant. The entire session was recorded (with the participant's camera off). Any scratch work made by the participant was collected electronically via pictures taken by each participant.

*Phase III: Technology Discussion*

During the first part of the interview, each participant revisited their ranks of five technology options from Phase I and discussed with the researcher about the pros and cons of each option and the rationale underlying the ranks. The purpose of this phase was to get participants acquainted with the various degrees of affordance and constraints that came with the different technology options.

*Phase IV: Statistical Thinking Assessment*

During this core part of the interview session, each participant was shown one warm-up interview question (item 1) and nine statistical problems (item 2 to item 10) with additional questions related to pedagogy (Appendix C). Topics covered both descriptive statistics and inferential statistics in a typical college-level introductory statistics course. Problems were selected from the LOCUS[1] assessment which emphasized four essential steps of statistical problem solving recommended by the GAISE report (Franklin et al., 2007).

[1]LOCUS is an NSF Funded DRK12 (DRL-1118168) project focused on developing assessments of statistical understanding.

Table 3.1. *Statistical Thinking Assessment in Statistical Problem Solving*

| Investigative Step | Question Item |
|---|---|
| Formulate Questions | 2 |
| Collect Data | 3 |
| Analyze Data | 4, 5, 6 |
| Interpret Results | 4, 7, 8, 9, 10 |

The pedagogical questions were created to provoke the different thinkings between statistics and mathematics. All participants verbally answered each problem and provided rationales, but some wrote down their problem-solving work. Additionally, they responded to questions that could reveal their thinking process of statistical problem-solving. For the sake of mimicking an actual testing environment, no correct answer was provided, nor participant's response was judged unless requested by the participant.

**Analysis**

Technology evaluation survey results and audio-taped interviews were transcribed and entered into a spreadsheet. The rows of the spreadsheet consisted of survey items and questions from all stages and phases. The responses were entered into columns for each participant. A separate spreadsheet was created that computed and summarized the mean scores and the standard deviation of each participant's evaluation in all domains of TPACK for the TPACK survey results.

To answer the first research question in relation to how statistical thinking and mathematical thinking take place among mathematics teachers when teaching introductory statistics, responses to statistical problems (item 2 to item 10 from the interview handout) in phase IV were compared and summarized among all participants. The main themes of differences between statistical thinking and mathematical thinking were constructed based on

the literature review of this study. They were coded into six main types of differences before data collection:

1. Problem-Solving Process
2. Variability
3. Context
4. Data Production
5. Transnumeration
6. Probabilistic Thinking

By employing grounded theory (Corbin & Strauss, 2014) along with the SKT framework's content knowledge (Groth, 2007) and LOTUS's grading rubric, mathematics teachers' ways of solving statistical problems were interpreted line-by-line, grouped by similar ideas, and labeled as mathematical thinking, statistical thinking, both, or neither based on the aforementioned six main themes of differences between statistical thinking and mathematical thinking. From these ideas, subthemes of differences between statistical thinking and mathematical thinking were also emerged and composed. These subthemes corresponded to some of the identified differences from existing literature and provided insight for new differences that other researchers did not mention. Specific responses from participants were discussed in detail as typical examples for each subtheme in which mathematics teachers addressed statistical problems. A colleague of the researcher who had coding experience went over the final coded subthemes and the corresponding examples to bring up new insights or potential biases. The final responses to the first research question were organized based on the four steps of an investigative study by the GAISE framework (Franklin et al., 2007): formulating questions, collecting data, analyzing data, and interpreting results.

To respond to the second research question regarding how mathematics teachers' thinking affects their teaching, responses to pedagogical questions (item 2 to item 10 from the interview handout) in phase IV were analyzed and categorized for each participant and across all participants with reference to the SKT framework's specialized knowledge (Groth, 2007). Mathematics teachers' pedagogical choices were reported for all participants. Mathematics teachers' rationale substantiated some pedagogical decisions. To address the effect of mathematics teachers' thinking on teaching, each participant's pedagogical choice, if there was any, was matched with their thinking based on the results from research question 1. Possible patterns of pedagogical choices were identified across all participants for each statistical concept. Effects were summarized in three areas: topic coverage in statistics, delivery methods in the class, and student assessments.

To address the third research question in regards to how mathematics teachers promote statistical learning, all responses in phase IV were analyzed for each participant and across all participants. In particular, by referring to the SKT framework (Groth, 2013) and studies that encourage exploratory and informal analysis in statistics (Bakker et al., 2006; Cobb & Moore, 1997; Gil & Ben-Zvi, 2011; Konold & Harradine, 2014; Makar, 2014), select mathematics teachers' pedagogical ideas were identified as pedagogically powerful ideas and KDU that can potentially transcend students' learning experiences in introductory statistics into something powerful and meaningful.

# Chapter 4

# RESULTS: PARTICIPANT STATISTICS

This study focused on the analysis of qualitative data from interviews. The results and analyses in this study were based on 15 mathematics teachers' responses to both nine statistical problems and pertinent pedagogical questions. For the first research question, participants' statistical solutions to item 2 to item 10 from the interview handout were designated as examples of mathematical thinking or statistical thinking. A framework was used to hypothesize aspects of mathematics teachers' thinking in four components of statistical problem solving: formulating questions, collecting data, analyzing data, and interpreting results. In the second research question, participants' pedagogical responses to item 2 to item 10 from the interview handout were analyzed and linked to their thinking regarding the first research question. Effects of mathematics teachers' thinking on their teaching of statistics were placed into three categories: topic coverage, delivery method, and student assessment. In the last research question, select mathematics teachers' responses to pedagogical questions in the interviews were extracted and unpacked as valuable examples of pedagogical content knowledge in statistics that could potentially enhance students' statistical understanding. When reporting results, participants were referred to by their randomly generated two-letter code name.

## Demographic Statistics

The majority of participants were females, but gender was not a factor considered in this study. The age distribution was not uniform, but both young and senior instructors were

Table 4.1. *Demographic Statistics*

| Gender | #Responses |
|---|---|
| Female | 9 (60.0%) |
| Male | 6 (40.0%) |

| Age | #Responses |
|---|---|
| 30-39 | 7 (46.7%) |
| 40-49 | 3 (20.0%) |
| 50 and above | 5 (33.3%) |

represented. Younger instructors offered more experience with modern technology and new experimental ideas in teaching, whereas senior instructors provided information based on their observation of changes which have occurred in statistics education over the past decade.

**Academic Background**

The college that the respondents were recruited from requires its mathematics instructors to possess a minimum of master's degree in a related field. The majority of the instructors hold a master's degree in mathematics or applied mathematics. At the time of the interview, a few of them were also pursuing a doctoral degree. Additionally, three participants had received professional training in education, with one related to statistics. The one participant who had only a bachelor's degree has been teaching at the college for a very long time. The median number of courses in statistics, mathematics, computer science, and

Table 4.2. *Academic Background*

| Highest Degree | #Responses |
|---|---|
| BS (Math) | 1 (6.7%) |
| MS (Applied Math) | 4 (26.7%) |
| MS (Math) | 3 (20.0%) |
| MS (Engineering) | 1 (6.7%) |
| EdD (Math Education) | 1 (6.7%) |
| PhD (Math Education) | 1 (6.7%) |
| PhD (Math/Statistics Education) | 1 (6.7%) |
| PhD (Math) | 3 (20.0%) |

| Undergrad GPA | #Responses |
|---|---|
| 3.00-3.49 | 7 (46.7%) |
| 3.50-3.99 | 8 (53.3%) |

| Graduate GPA | #Responses |
|---|---|
| 3.50-3.99 | 14 (93.3%) |

education taken by all participants were 1, 20, 1, and 0, respectively, with interquartile ranges (IQR) of 1.5, 8.5, 2, and 3, correspondingly. The median length of teaching experience in introductory statistics was 5 years with an IQR of 7 years. These 15 participants formed a group of mathematics teachers who had sufficient education in mathematics and good experience in teaching statistics but insufficient training in statistics and education. This unique group characteristic fostered the detection of the differences between their mathematical thinking and statistical thinking in their approaches to teaching statistics.

### TPACK Knowledge

During the online TPACK survey, all participants rated themselves on seven domains of the TPACK framework. Table 4.3 shows the mean and the standard deviation for each score. To identify scores that reflect opinions that are far from neutral, the last column of the table gives the relative difference between the mean score and a neutral score of 3 for each domain of knowledge:

$$\text{Relative Score} = \frac{\text{Mean} - 3}{\text{Std Dev}}.$$

According to the table, the relative score for CK, PK, and PCK are much higher than the rest, suggesting that the mean score in these domains could be very different from the neutral score. The 15 mathematics teachers rated themselves highly in their content knowledge, pedagogical knowledge, and pedagogical content knowledge. Since a score of 3 could also be used when the statement in the survey was not applicable, no meaningful information might be extracted from teachers' knowledge in other domains. However, it was noticeable that these mathematics teachers were not confident in their technological knowledge in general.

Table 4.3. *TPACK Statistics*

| Knowledge | Mean | Std Dev | Relative Score |
|---|---|---|---|
| TK | 2.51 | 1.29 | −0.38 |
| CK | 4.24 | 0.39 | 3.21 |
| PK | 4.26 | 0.52 | 2.44 |
| PCK | 3.93 | 0.70 | 1.33 |
| TCK | 3.27 | 1.22 | 0.22 |
| TPK | 3.08 | 1.02 | 0.08 |
| TPACK | 3.27 | 0.80 | 0.33 |
| Overall | 3.42 | 0.61 | 0.69 |

Because it is a self-reported evaluation completed by mathematics teachers teaching statistics, the potentially significantly high scores in content knowledge and pedagogical knowledge could be teachers' over-estimation. However, those scores could also be interpreted as signaling that the participating mathematics teachers were very confident about their content knowledge and pedagogical knowledge in introductory statistics.

The online survey also looked into participants' knowledge related to technology in the form of free response questions. The mandatory technology usage made by participants' college was almost identical for all mathematics teachers. Most teachers claimed that the statistics course they were teaching required students to use a scientific calculator. Moreover, more than half of the interviews took place during the pandemic in 2020. As a result, many mathematics teachers also mentioned using the video conference software Zoom or Blackboard Collaborate Ultra.

Second, 11 mathematics teachers reported that they had implemented technology in their statistics course. Besides using learning management systems such as Blackboard, Canvas, or Brightspace, the most popular choice of technology among all participants was Excel for in-class demonstration and class projects. Other technology options, such as R and

web applets, were also mentioned as being used for similar purposes. Additionally, one teacher used DataCamp for data visualization and projects, one teacher took attendance in Google Sheets, and one teacher used Clickers to give pop quizzes in class. On a scale of 5, where 5 represents a highly positive experience in using technology to teach statistics, 27.2% (3/11) gave their experience of technology the highest rating, 18.2% (2/11) rated their experience 4, and 54.5% (6/11) rated their experience 3. According to the 11 mathematics teachers, technology enriches students' learning experiences in statistics by simplifying calculations, presenting data visually, and simulating statistical scenarios. Additionally, technology engages students in class and reduces students' anxiety on tests. However, many of them also stated that they needed more guidance and training in implementing educational technology effectively and creatively. One mathematics teacher expressed his concern that some technology options could be costly.

Among the four mathematics teachers who had not implemented any technology in their statistics course, three reasoned that they did not have sufficient knowledge in educational technology. One stated that technology implementation was not required for his statistics course.

Third, regardless of participants' experience in statistical technology implementation, all mathematics teachers participated in a technology evaluation survey in which they explored different types of technology options.

Table 4.4. *Five Technology Options: Familiarity*

| Option | #(Heard about it) | #(Used it) |
|---|---|---|
| Minitab | 6 (40.0%) | 0 (0.0%) |
| Tinkerplots | 3 (20.0%) | 1 (6.7%) |
| Google Sheets | 13 (86.7%) | 10 (66.7%) |
| Web Applets | 8 (53.3%) | 6 (40.0%) |
| Multimedia Product | 5 (33.3%) | 4 (26.7%) |

Based on Table 4.4, Google Sheets was the most familiar technology option among all mathematics teachers. Additionally, many mathematics teachers had either seen or used web applets for statistics. Surprisingly, the educational software Tinkerplots, which educators and researchers highly appraised, was unfortunately unrecognized by most participating mathematics teachers.

After free exploration, mathematics teachers ranked five technology options based on whether or not they would implement them in their teaching of statistics. Table 4.5 summarized the overall ranking among the 15 mathematics teachers. If an option was ranked the first or the last (no. 6), its frequency of occurrence was indicated in parentheses. The last column computed the relative rank compared to a middle rank of 3:

$$\text{Relative Rank} = \frac{\text{Mean Rank} - 3}{\text{Std Dev}}.$$

Note that, in addition to five technology options, there was an extra option–"none"–that participants could choose in their ranking. This option was used to separate options that participants would implement from those they would not. In extreme cases, "none" could even be ranked first if participants decided to not implement any of the five technology options. Also, participants were allowed to assign the same rank to multiple options. In the final responses collected, all mathematics teachers except one assigned a different rank to each technology option.

According to Table 4.5, the most favorable option was the Rice Virtual Lab in Statistics, ranked first eight times by the respondents. The following reasons were given:

1. It's a complete package for learning statistics. (5/8)

Table 4.5. *Five Technology Options: Rank*

| Option | Mean Rank | Highest Rank | Lowest Rank | Relative Rank |
|---|---|---|---|---|
| Minitab | 3.33 | 1 (×2) | 6 (×1) | 0.22 |
| Tinkerplots | 3.87 | 2 | 6 (×2) | 0.70 |
| Google Sheets | 3.80 | 1 (×2) | 6 (×3) | 0.44 |
| R/C Applets | 3.13 | 1 (×2) | 6 (×2) | 0.08 |
| Rice Virtual Lab | 2.20 | 1 (×8) | 5 | −0.50 |
| None | 4.53 | 1 (×1) | 6 (×6) | 0.93 |

2. It's free. (4/8)
3. It's ready to use. (2/8)
4. The participant had a good experience using it before. (1/8)

Mathematics teachers who did not assign a high rank to the Rice Virtual Lab argued that:

1. It's hard to navigate through different links and pages. (2/5)
2. Its content is not customizable or modifiable. (2/5)
3. Its content is not well organized. (1/5)
4. It's more suitable for self-paced learning. (1/5)

Based on rationales provided for the higher rankings, many of the mathematics teachers preferred technology options that were ready to implement with customizable learning materials, such as applets, lesson plans, worksheets, and datasets. Ideally, the option could be used for teaching more than one topic in introductory statistics. They would also prefer that the option was free of charge or provided discounts for students and educators.

Tinkerplots was rated as the least favorable technology option by all of the mathematics teachers. It appeared to them that Tinkerplots was designed mainly for elementary and secondary statistics education. They worried that it might not be sufficient for college-level introductory statistics courses. Additionally, Tinkerplots looked more complicated than other

options in that it might expect a steeper user learning curve. However, some mathematics teachers who rated Tinkerplots higher held different opinions. They alleged that Tinkerplots had a unique functionality that lets users interact with statistics and graphs, which was much better than staring at static outputs produced by other software such as Minitab. Even though some web applets also allowed user interaction, they usually lacked user customization. In Tinkerplots, teachers could design and modify the program to meet specific teaching needs. In some respondents' opinion, Tinkerplots could be made sophisticated enough to accommodate college-level introductory statistics.

### Statistical Response Accuracy Map

All participants' responses to nine statistical problems (item 2 to item 10 from the interview handout) were graded based on the LOTUS's grading rubric (see Appendix C). Each part of a problem was graded individually. If a response was correct, partially correct, or incorrect, it received 1, 0.5, and 0 points, respectively. In item 7, parts 2 to 4 were not graded because these were open-ended questions. The maximum number of points a participant could receive was 17. Table 4.6 summarized all participants' accuracy results. Correct, partially

Table 4.6. *Statistical Response Accuracy Map*

| ID | % | Point | 2 | 3 | 4 | | | 5 | 6 | 7 | | 8 | | 9 | | | 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HH | 97.1% | 16.5 | | | | | | | | | | | | | | | | | |
| OZ | 91.2% | 15.5 | | | | | | | | | | | | | | | | | |
| RM | 88.2% | 15 | | | | | | | | | | | | | | | | | |
| KU | 88.2% | 15 | | | | | | | | | | | | | | | | | |
| DE | 85.3% | 14.5 | | | | | | | | | | | | | | | | | |
| WG | 76.5% | 13 | | | | | | | | | | | | | | | | | |
| AQ | 76.5% | 13 | | | | | | | | | | | | | | | | | |
| PK | 73.5% | 12.5 | | | | | | | | | | | | | | | | | |
| PV | 73.5% | 12.5 | | | | | | | | | | | | | | | | | |
| FX | 73.5% | 12.5 | | | | | | | | | | | | | | | | | |
| MP | 70.6% | 12 | | | | | | | | | | | | | | | | | |
| RY | 67.6% | 11.5 | | | | | | | | | | | | | | | | | |
| QK | 64.7% | 11 | | | | | | | | | | | | | | | | | |
| YW | 52.9% | 9 | | | | | | | | | | | | | | | | | |
| GG | 38.2% | 6.5 | | | | | | | | | | | | | | | | | |
| | | | 73.3% | 93.3% | 100% | 100% | 26.7% | 40.0% | 93.3% | 53.3% | 73.3% | 0.0% | 100% | 13.3% | 33.3% | 100% | 46.7% | 46.7% | 66.7% |

correct, and incorrect responses were colored in green, yellow, and red, respectively. The table was ordered by accuracy in descending order. The last row of the table indicated the percentage of entirely correct responses to each problem. For a more detailed accuracy map that included additional coded responses, see Appendix D.

Some striking observations could be made based on the accuracy map. Among the top three mathematics teachers, it was not so surprising to see HH and RM since they were the only two participants who reported to have received some formal training in statistics. OZ, who possessed the lowest academic degree level out of all 15 teachers, demonstrated her solid statistical understanding regardless of her lack of formal statistics education. These understandings might have been gradually developed over her 15 years of teaching statistics. GG, who claimed having taken eight statistics courses, shockingly did not do well compared to other participants who took fewer statistics courses.

# Chapter 5

# RESULTS: RESEARCH QUESTION 1

This chapter addressed the first research question: *In which ways do statistical thinking and mathematical thinking take place among mathematics teachers when teaching introductory statistics?* In each section, participants' responses for each corresponding statistical problem (item 2 to item 10 from the interview handout) were reported, analyzed, and categorized based on the common knowledge from the SKT framework I (Groth, 2007). To summarize findings, the researcher created a coding framework (see section Hypothesized Aspects of Teachers' Thinking in Statistics on page 113). This framework served as sub-themes that emerged from the six major themes identified by the literature review: problem-solving process, variability, context, data production, transnumeration, and probabilistic thinking. More importantly, the framework reflected prominent differences identified by all participants between statistical thinking and mathematical thinking across four statistical problem-solving components: formulating questions, collecting data, analyzing data, and interpreting results. By referring to the coding framework, select responses to nine statistical problems from all 15 participants were discussed and labeled as examples of statistical thinking or mathematical thinking.

## Item 2: Statistical Problem

The first statistical problem during the interview involved choosing a statistical question that could be addressed using a specified sample.

> For their final project, students in a math class are required to answer a question by collecting data about students at their school. For which of the following questions could a random sample of students provide the best approximate answer?

(A) How many students attend the school?
(B) How many hours does each class at the school meet per year?
(C) How many text messages do students at the school send per week?
(D) Do students at this school have higher test scores than students in other schools in the district?

**Responses**

Out of 15 participants, 11 selected the correct answer option C only. Among the four who chose incorrectly, three provided multiple options, including the correct option C. The most popular incorrect selection was option D. Even though this question did not require participants to perform any computation, many participants spent a substantial amount of time on this problem. Some commented that they had never seen or used a similar problem in their teaching career of statistics.

Table 5.1. *Responses to Item 2*

| Option | #Responses |
|---|---|
| A | 0 |
| B | 2 |
| C | 14 |
| D | 3 |

Participant QK explained that option C sounded "more like a textbook question" that a random sample could answer. WG, PK, and HH talked about the process, including using a confidence interval to estimate the population's mean number of texts sent per week. Additionally, DE and QK explicitly mentioned that option C evoked variability in students' various responses. The following are the reasons given for not choosing option A:

1. You cannot answer the question with a random sample. (10/15: WG, RY, QK, GG, PK, KU, HH, AQ, FX, YW)

2. The total number of students attending the school when the question is being asked is deterministic and not worth studying. (2/15: OZ, MP)
3. We are collecting information from a sample of students. The total number of students attending the school is not a parameter of students, so it cannot be estimated using our sample. (2/15: DE, RM)
4. Students usually do not know how many students attend the school. So it is impossible to obtain useful data that can answer the question. (1/15: PV)

The following are the reasons given for not choosing option B:

1. This question can be answered by a random sample, but it is harder than C. (5/15: QK, PK, KU, AQ, YW)
2. The number of class hours per week is deterministic and dictated by the class when the question is being asked; it is not worth studying. (3/15: PV, OZ, MP)
3. We are collecting information from a sample of students. The number of class hours per week is not a parameter of students, so our sample cannot estimate it. (2/15: DE, RM)
4. This question is confusing to me (i.e. What does class mean?). I do not know how to answer that by a random sample. (2/15: RY, HH)
5. This question requires the collection of more than one sample. (1/15: FX)

The following are the reasons given for not choosing option D:

1. This question requires the collection of more than one sample (from other schools). (11/15: PV, DE, WG, RY, QK, RM, PK, KU, HH, FX, OZ)
2. This question can be answered by a random sample, but it is harder than C. (1/15: AQ)

**Mathematical Thinking**

In mathematics, students usually do not need to come up with questions themselves. Question formulation is unique to statistics, but formulating questions that demand a deterministic answer is considered mathematical. For instance, OZ stated that the answers to options A and B were deterministic and not worth studying. Though more than half of the participants did not provide an explicit reason for not choosing option A, nor did they understand completely what statistical questions were, they were all very affirmative that option A could not be answered by a random sample. Some participants suggested obtaining the answer from the office of the registrar.

Additionally, three participants brought up the mathematical recipe of applying the confidence interval to estimate the population mean:

> WG: We can ask students how many texts they sent last week and then use a confidence interval to estimate.
>
> PK: A confidence interval for mean can be created to estimate the mean text messages sent by students at that school per week.
>
> HH: This question can be answered by a confidence interval estimate of the population mean texts sent by students per week based on a simple random sample which gives us a lower bound and an upper bound. Depending on how good the sample is and the sample size, we may have a pretty narrow interval and a good estimate.

Because these participants taught statistics in the past, they knew the bigger picture in statistical problem-solving; they knew what happens after formulating such a question and collecting data. Nonetheless, these responses were also labeled as statistical thinking because they understood which statistical method should be implemented to address the question.

In addition to the mathematical process of determining a confidence interval, WG mentioned a survey question to collect data. His question specified the time frame of a

particular week–last week–leading to a more deterministic answer from respondents. Additionally, HH's responses suggested that people who used mathematical thinking tended to be result-oriented, preferring a narrower interval for mathematically higher precision. However, in statistics, even a narrow confidence interval could be totally off in capturing the true parameter and thus should be considered as equally as wider ones.

**Statistical Thinking**

There were two types of statistical thinking involved in this problem. The first one consisted of recognizing that a statistical question anticipated variability in responses collected and accounted for the variability when interpreting results. For option C, one could use WG's survey question, "how many text messages did you send last week?", to collect responses. One anticipated a list of different numbers of text messages sent from sampled students. Moreover, the estimated parameter might even vary depending on the time the question was being asked. On the other hand, if the knowledge of the total number of students attending the school was widely known to all students, then a survey question such as "what's the total number of students attending this school this semester?" would possibly produce a list of identical or similar responses from sampled students. The estimated parameter here would not change much if the period was limited to be a specific semester. Even though a few participants vaguely mentioned or implied the meaning of a statistical question, only two participants explicitly articulated the anticipation of variability in responses:

> DE: C is a question that expects different responses from surveyed subjects, or what statistics calls variability. It's highly possible that they will report different numbers of text messages sent per week.
>
> QK: It sounds more like a statistical question to me. We did have a brief discussion at the beginning of the semester on statistical questions and survey questions. C is

> a typical statistical question that could result in variability in the data collected–in this case, the number of text messages sent per week.

DE demonstrated his understanding of the difference between a deterministic and a stochastic question. Among the remaining mathematics teachers, a lack of statistical thinking may not be too surprising, considering that more than half (10/15) of the participants claimed that they did not introduce the concept of statistical questions at all.

The second type of statistical thinking involved understanding that statistics could be used to gain information about a population. However, it required defining the correct population and collecting a sample from that population. Inferences made from that sample were only valid if the sample chosen was representative, and a random sampling method tended to yield a more representative sample. Almost all participants emphasized "a random sample" in their rationales. But not all participants managed to identify the correct population to be sampled from for each option. In fact, there was a quick way to answer this statistical problem about formulating statistical questions–checking the population of interest. The problem pointed out that we had a random sample of students at their school, which determined the population of interest: students at that school only. Thus, any statistical questions raised must be about some characteristics of students at that school only. DE and RM recognized that options A and B sought characteristics of the school instead of students, whereas 11 participants acknowledged that option D compared characteristics of students of that school to characteristics of students from other schools in the district. Only option C studied aspects of the desired population. According to the participants above' elimination process of options A, B, and D, it was easier for them to recognize the population of interest when comparing characteristics of two groups of subjects. Even though many failed to see the

different populations of interest in options A and B, participants knew that something was not right in those options–they agreed that options A and B were harder to address by a random sample of students than option C. But they could not articulate why precisely.

### Item 3: Statistical Problem

The second statistical problem tested participants' ability to select a representative sample:

> A student wants to estimate the mean number of books that have been read by all students at his school over the summer. On Monday morning, he will survey the first 35 students who enter the library. Is this the best way to select a sample for this purpose?
> (A) No. The student should survey students entering the library on more than one day of the week.
> (B) No. The student should take a random sample of students entering the library instead.
> (C) No. The student should take a random sample of students from all students, not just those entering the library.
> (D) Yes. Selecting a sample in this way will not introduce the possibility of bias.

**Responses**

Thirteen participants selected option C as their first and final answer. Participant AQ chose option B first but switched to option C later. Participant YW selected option B.

Table 5.2. *Responses to Item 3*

| Option | #Responses |
|---|---|
| A | 0 |
| B | 1 |
| C | 14 |
| D | 0 |

All participants agreed that the sample initially chosen by the student was biased. Surveying the first 35 students who entered the library was considered convenience sampling, which was easy to do but not random since not all students at his school had an equal chance

of being surveyed. No participants chose option D. The student wanted to estimate the mean number of books read by all students at his school, suggesting that the population of interest should include all students at his school. Option A surveyed students entering the library on more than one day of the week. It was an improvement over the original method, but it excluded students who did not enter the library. Option B emphasized the notion of "random sample" and appeared to be sound. However, its population of interest was still limited to students who entered the library only. Option C extended option B to include all students at the school, which guaranteed that everyone in the population of interest had an equal chance of reporting their number of books read, and this was the best means of sampling.

**Mathematical Thinking**

When explaining why she thought option B was the correct answer, YW stated that,

> I picked B because the sample should be a subset of the population, and [the group of] students entering the library is a subset of all students at that school. Also, we should always make sure the sample is random so that it's fair to everyone in the population and the results are more accurate.

First, the word "subset" is borrowed from mathematics, yet it is widely used by many statistics textbooks. Many students taking introductory statistics courses may not have a background in mathematics. Thus they may lack the appropriate understanding of the word "subset" and what this word entails about the relation between a sample and a population. Even if they understand what "subset" means, this word does not reveal the interconnection between a sample and a population; a good sample should represent the population and students who do not understand that will have difficulty grasping sampling variability. Second, YW demonstrated her additive view of a sample by stating that "[the group of] students entering the library is a subset of all students at that school." It is a true statement if we look at it alone.

But when we include the context of the entire problem, such a sample is insufficient to estimate the mean number of books read by all students at the school. By specifying a random sample, YW will only be able to estimate the mean number of books that have been read by all students who enter the library. Lastly, YW's wording choices such as "fair" and "accurate" suggested that YW emphasized obtaining a more theoretically-acceptable answer with minimal numerical error in her statistical problem-solving.

**Statistical Thinking**

All participants acknowledged the importance of a representative sample for validating inferences about a population. Unlike in item 2, almost all participants managed to identify the correct population of interest. However, RY's rationale went even further:

> The best way to select a sample for this purpose should be a way that selects a sample that includes as many relevant characteristics as you can from the original population. In this case, since the population is all students at his school and the parameter of interest is the mean number of books that have been read, it makes sense that we should consider both groups of students who go to the library and [those] who don't. Students who go to the library will probably have a higher mean number of books. I think a good sample is one that includes students who go to the library–who probably like to read books but could also be just going there for a quick nap, you know, as well as students who rarely or never enter the library–who probably don't like to read books, which represents same characteristics of the original population. Such a sample will probably produce a result that's more compelling.

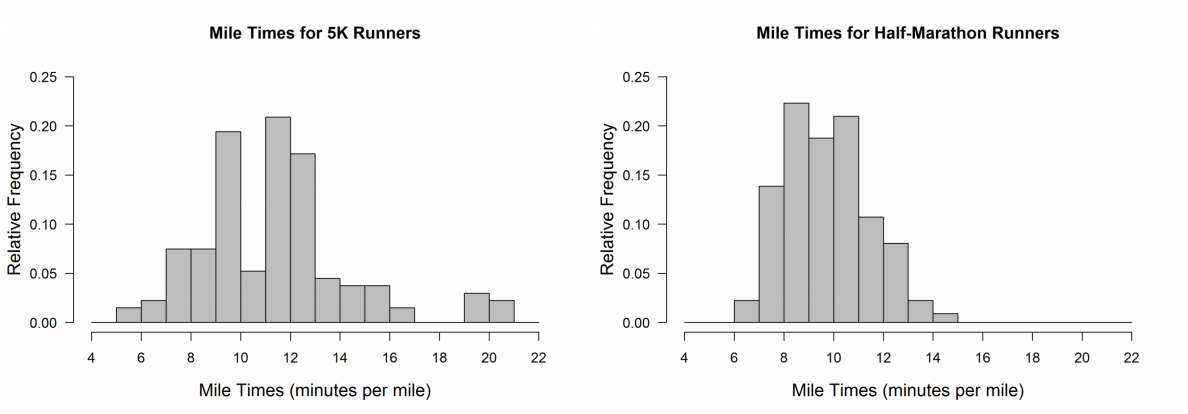
RY demonstrated a multiplicative view of a sample and went beyond what other participants stated. According to RY, a good sample was not only a part of the population but also one that represented various characteristics of the population well–"a mini version of the population." Moreover, RY understood that statistics looked for a practical solution that told stories behind numbers.

### Item 4: Statistical Problem

The third statistical problem tested participants' ability to compare descriptive statistics from two histograms:

> The city of Gainesville hosted two races last year on New Year's Day. Individual runners chose to run either a 5K (3.1 miles) or a half-marathon (13.1 miles). One hundred thirty-four people ran in the 5K, and 224 people ran the half-marathon. The mile time, which is the average amount of time it takes a runner to run a mile, was calculated for each runner by dividing the time it took the runner to finish the race by the length of the race. The histograms (see Figure 5.1) show the distributions of mile times (in minutes per mile) for the runners in the two races.

Figure 5.1. *Item 4: Center, Variability and Outliers*




1. Jaron predicted that the mile times of runners in the 5K race would be more consistent than the mile times of runners in the half-marathon. Do these data support Jaron's statement? Explain why or why not.
2. Sierra predicted that, on average, the mile time for runners of the half-marathon would be greater than the mile time for runners of the 5K race. Do these data support Sierra's statement? Explain why or why not.
3. Recall that individual runners chose to run only one of the two races. Based on these data, is it reasonable to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K? Explain why or why not.

**Responses to Question 1**

The first question asked participants to compare the variability of the two distributions. All participants disagreed with Jaron's statement, claiming that mile times for half-marathon runners (second histogram) were more consistent with less variability. Participants' justifications were primarily based on three aspects:

1. Comparing the shape of two histograms, the second histogram looks more narrow and thus more consistent. Participants used the following words to describe the shape of two distributions: "spread" (8/15: PV, WG, QK, GG, RM, HH, FX, MP), "clustered" (1/15: FX), and "narrow" (1/15: AQ).
2. Comparing the variability of two histograms, the second histogram suggests smaller measures of variation, i.e., range, standard deviation. In addition to the word "consistent" from the problem, participants referred to variability using the following words: "variation" (3/15: PK, YW, OZ), "dispersion" (1/15: KU), and "similarly" (1/15: DE).
3. The first histogram contains noticeable outliers to the right that increases the variability of the distribution (1/15: FX).

**Responses to Question 2**

The second question asked participants to compare the mean of the two distributions. All participants disagreed with Sierra's prediction and claimed that mile times for 5K runners (first histogram) had a higher average. Interestingly, in all 15 responses, participants applied different terms to describe "average":

1. The average of the first histogram looks greater. (6/15: PV, WG, QK, RM, PK, FX)

2. The average of the first histogram looks greater due to the outliers to the right. (4/15: RY, AQ, YW, MP)
3. The mean of the first histogram looks greater due to the outliers to the right. (3/15: GG, HH, OZ)
4. The mean of the first histogram looks greater. (1/15: KU)
5. The median of the first histogram looks greater. (1/15: DE)

Five participants assigned a numerical estimate to the "average" of each histogram and made their response. A few argued without estimation, reasoning that the presence of large values in the first histogram would result in a higher "average." Needless to say, most teachers were aware of the need to take the variability into account when comparing the mean of the two distributions.

**Responses to Question 3**

Eleven participants answered "yes." However, four participants (DE, RM, HH, OZ) answered "no" which was the correct answer. For instance,

> OZ: No. These histograms are composed of two groups of different runners. It is reasonable to assume so if these histograms are composed of two groups of the same runners.
>
> HH: No. Because these two histograms are not based on the same individual running both marathons. So they cannot be used to answer the question. This question is tricky. It is very tempting to answer yes, but the first sentence, which says that "individual runners chose to run only one of the two races," generously provides an important hint that the population must be the same for both datasets.

Even though participant KU answered this part incorrectly, she did notice that the first sentence of question 3 seemed "out of place." KU commented that "I feel like I didn't use the information from the first sentence. Did I do something wrong?"

**Mathematical Thinking**

When addressing question 2, 10 participants used the word "average" from the problem. They did not specify which measures of the center they were comparing. In mathematics, the word "average" usually refers to the arithmetic mean. So it is not wrong to use the word "average" directly. But, in statistics, the word "average" is not used. In some textbooks, "average" represents the arithmetic mean but is called "mean" only. In some other textbooks, "average" is a general term for measures of central tendency and can represent mean, median, or mode. It is often recommended by statistics textbooks that statistics instructors should use "mean" in place of the more mathematical "average" to avoid confusion and reinforce statistical vocabulary among students.

In question 3, 11 participants revisited their responses to question 2 and connected question 3 and question 2. Such a way of thinking reflected strongly mathematical thinking. Under such thinking, the description in question 3 was over-simplified to comparing means of two distributions only, which was the task in question 2. As a matter of fact, these mathematics teachers omitted a vital aspect of statistics–data production–which is essential in selecting a statistical method or validating a selected statistical model. To justify the given conclusion in question 3, one needs to resample and recollect data on the same group of individuals twice.

**Statistical Thinking**

Many participants demonstrated their statistical ability to estimate and compare descriptive statistics from two histograms. Based on responses to questions 1 and 2, all participants visually compared the center, variability, and overall shape of two distributions.

Many demonstrated their knowledge about the effect of outliers on measures of center. For example,

> HH: Due to the right skewness of the 5K marathon, its mean is pulled over to the right and greater than the mean of the half-marathon's.
>
> OZ: The first histogram presents several extremely high values, shifting the mean to the right.

They used this knowledge to compare the means of the two histograms without finding out the numerical means. At a higher level, when addressing the last question, four participants were able to question how data were produced and concluded that the current data was insufficient to justify the conclusion that the mile time of a person would be less when that person ran a half-marathon than when he or she ran a 5K.

**Item 5: Statistical Problem**

The fourth statistical problem tested participants' ability to compare consistency in categorical data.

> A school is planning a field trip to the aquarium or to the zoo for students in grades 6 through 9. There are 100 students in each grade level and every student was asked which place he or she would prefer to visit. The bar graphs for the four grade levels are constructed (see Figure 5.2). In which grade level were the responses most consistent?
>
> (A) Grade 6 (B) Grade 7 (C) Grade 8 (D) Grade 9

**Responses**

Among the nine participants who selected grade 7, all of them argued that the heights of the bar for the aquarium and the bar for the zoo were the same for grade 7. Since there was no difference, the responses of grade 7 were more consistent.

Figure 5.2. *Item 5: Variability*

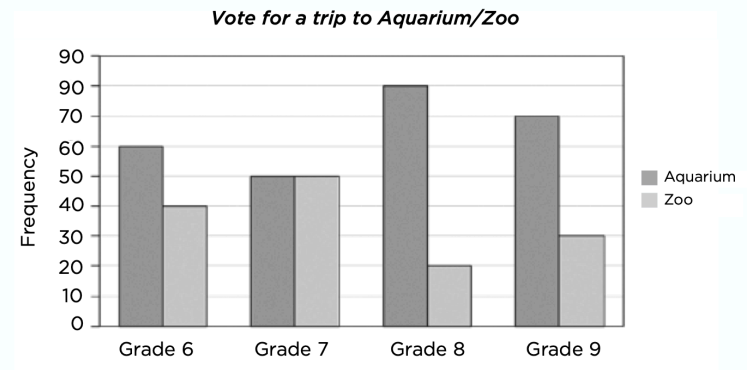



Table 5.3. *Responses to Item 5*

| Option | #Responses |
|---|---|
| A | 0 |
| B | 9 |
| C | 6 |
| D | 0 |

Among the six participants who selected grade 8 correctly, there were mainly two types of explanation:

> KU: In grade 8, we see that 90 students picked the aquarium and 20 picked the zoo. The option aquarium is chosen by 90 divided by 110, which is . . . more than 80% of all 8th graders. Looking at other grades, 60% of 6th graders picked aquarium, 50% [of] 7th graders picked aquarium, and 70% [of] 9th graders picked aquarium. Apparently, 8th graders have much more votes for the aquarium and fewer votes for the zoo–the responses are more consistent.
>
> HH: If we measure the visual difference between the frequencies of the aquarium and the zoo for each grade, grade 8 has the largest difference, and grade 7 has no difference, so the correct answer must be one of them. We also know that responses are either the aquarium or the zoo–they are qualitative. The bigger the difference in frequency, the less the variation, and thus more consistent. So, grade 8 is the correct answer.

Interestingly, the figure for the problem had a typo where on the vertical axis, there were two frequencies of 90–one of them should be 80. However, none of the participants noticed it.

Many participants focused their rationales only on the actual bars themselves. For those who

used relative frequency to argue, as KU did, they all used 90/110 as the basis of their response with no doubts.

**Mathematical Thinking**

Visually, the two bars in grade 7 have the most consistent heights. Without too much thinking, grade 7 seems to be the first choice. But that is true if only the heights or the frequencies of the bars for the aquarium and the zoo are considered. When taking the context of the problem into account, 7th graders did not obtain meaningful voting results since each option was picked by exactly half of the students. In fact, 7th graders have the least consistency in their responses.

**Statistical Thinking**

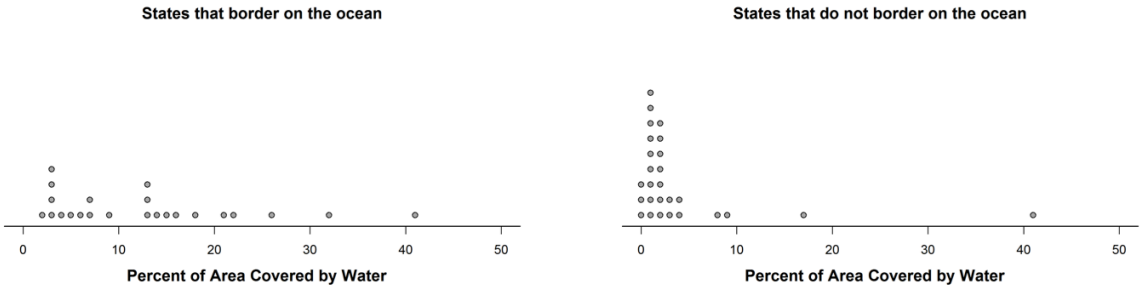
Many participants mixed responses up with frequencies when discussing the consistency. Those nine participants who selected grade 7 addressed consistency in frequency, not consistency in responses. It is ubiquitous for instructors and students to confuse frequencies in a frequency table or a bar graph with the original data values. High consistency in students' responses between the aquarium and the zoo means that most students picked the same choice. KU and HH demonstrated their understanding of how the given bar graph was produced from data and how they could be transformed back to data, which was essential to answer the problem correctly.

### Item 6: Statistical Problem

The fifth statistical problem tested participants' knowledge about the effects of outliers on descriptive statistics.

Carlton found data on the percent of area that is covered by water for each of the 50 states in the U.S. He made two dotplots (see Figure 5.3) to compare the distributions for states that border an ocean and states that do not border an ocean. Which of the following is the best statistical reason for using the median and interquartile range (IQR), rather than the mean and standard deviation, to compare the centers and spreads of these distributions?

Figure 5.3. *Item 6: Outlier-Resistant Measurements*




(A) The mean and standard deviation are more strongly influenced by outliers than the median and IQR.
(B) The median and IQR are easier to calculate than the mean and standard deviation.
(C) The two groups contain different numbers of states, so the standard deviation is not appropriate.
(D) The two distributions have the same shape.

**Responses**

Except for participant GG, all other participants selected the correct answer option A. GG's rationale was that when computing IQR, one needed to find the median. So it was easy

Table 5.4. *Responses to Item 6*

| Option | #Responses |
|---|---|
| A | 14 |
| B | 1 |
| C | 0 |
| D | 0 |

to find both. She claimed that the mean and standard deviation were irrelevant when finding the median and IQR. It appeared that GG misunderstood the problem and what it was asking.

Among the rest of the participants, there were primarily three ways to justify the selection of option A:

1. The presence of extreme values, such as outliers, makes the mean and standard deviation biased. The median and IQR are more resistant to the addition of a few extreme values because of how they are computed mathematically. (9/15: PV, DE, RY, HH, AQ, FX, YW, OZ, MP)
2. When the distribution of the data is asymmetrical, the median and IQR are preferred. Skewness influences the mean and standard deviation more than it does the median and IQR. (2/15: RM, KU)
3. The other three options are incorrect. (3/15: WG, QK, PK)

Some participants illustrated how the effect of outliers occurred in this problem. For instance,

> OZ: There are outliers in both dotplots. We know that outliers will greatly increase the mean and standard deviation, so the mean and standard deviation become biased. The median and IQR are less affected, and so it's better to use them to describe the data.
>
> MP: The calculation of the mean includes all data values, but the calculation of the median only uses at most the middle two numbers in the [ordered] data, so the mean is more influenced by outliers. The standard deviation depends on the mean, so it's also more influenced by outliers. The IQR uses Q1 and Q3, which are less likely to be outliers, so it's less influenced by outliers.

In OZ's description, the result of outliers' effect was specified–that both the mean and standard deviation will be higher. In MP's description, the mechanism of outliers' effect was explained.

The following are the reasons given for not choosing option B:

1. Easier calculation is not a statistical reason. (11/15: PV, WG, RY, QK, RM, PK, KU, HH, AQ, FX, OZ)

2. It does not explain the effect of outliers on measures of center and variation. (2/15: YW, MP)
3. If technology is used, descriptive statistics are equally easy to calculate. (1/15: WG)
4. Calculating the median is easier, but calculating the IQR is not easy as it uses the percentile formula. (1/15: DE)

The following are the reasons given for not choosing option C:

1. The standard deviation can be used to compare groups with different sample sizes because its calculation accounts for the sample size $n$. (7/15: DE, WG, RM, KU, HH, FX, OZ)
2. It does not explain the effect of outliers on measures of center and variation. (2/15: YW, MP)
3. The standard deviation can be used to compare groups with different sample sizes because it is similar to the mean absolute difference. (1/15: RY)
4. The standard deviation can be used to compare groups with different sample sizes if the means of two groups are not too different. (1/15: PK)
5. The sample sizes looks similar for the two groups. So it's ok. (1/15: QK)
6. The sample size is irrelevant. (1/15: PV)
7. The option does not mention the mean. (1/15: AQ)

The following are the reasons given for not choosing option D:

1. The two distributions do not have the same shape. (7/15: PV, DE, WG, RY, QK, RM, PK)
2. Having the same shape is irrelevant. (3/15: HH, AQ, OZ)
3. Having the same shape is true but irrelevant. (2/15: KU, FX)

4. It does not explain the effect of outliers on measures of center and variation. (2/15: YW, MP)

**Mathematical Thinking**

When explaining why the standard deviation could compare groups with different sample sizes, the participants gave three reasons. The first reason, mentioned by seven participants, came from the mathematical formula for the standard deviation for which the sample size $n$ was accounted. The second reason, given by participant RY, connected the standard deviation to the mean absolute difference which was a type of mean. If the mean could be used to compare groups of various sizes, so could the standard deviation. This reason demonstrated RY's horizon knowledge in making connections between different statistical concepts. The third reason, given by participant PK, tackled the question at a different angle; he claimed that the mean for the two groups looked similar in the dotplots, so it was reasonable to use the standard deviation to compare them. PK remembered examples from a textbook in which variability was compared between two groups of different sample sizes and very distinctive sample means. The textbook used the coefficient of variation (CV) which was also known as the relative standard deviation (dividing standard deviation by its mean). The CV normalized the standard deviations from two different groups based on their means to be comparable. Even though the third reason given here did not explain directly why the standard deviation could be used to compare groups with different sample sizes, it addressed another potential question often raised by statistics learners: can the standard deviation be used to compare groups with very different means?

**Statistical Thinking**

The most crucial statistical thinking from this problem is understanding how the outliers affected descriptive statistics. While all participants, except GG, agreed on using the median and IQR when outliers were present, only OZ and MP specifically explained what the effect was and why this effect took place. OZ, who had taught statistics for 15 years, told the researcher that her students would have no problem writing a statement, such as option A, but that did not mean that they understood it. Her students could be reciting rules from the textbook or lecture notes. In her opinion, only when students explained the impact within the specific context of the problem would they demonstrate their understanding of the effect of outliers on descriptive statistics.

**Item 7: Statistical Problem**

The sixth statistical problem tested participants' knowledge of probability.

> Question 1: Assume a coin is fair. If we toss the coin five times, how many heads will we get?
>
> Question 2: You pick up a coin. Is this a fair coin?
>
> 1. Provide answers to each question.
> 2. How are these two questions similar?
> 3. How are these two questions different?
> 4. Is there any relationship between the concepts underlying these two problems? If so, what is it?

**Responses to Part 1**

For question 1 regarding the number of heads when tossing a coin five times, 13 of 15 participants recognized that the number of heads was stochastic. Among these 13 participants, seven of them correctly gave all possible numbers of heads. Among these seven participants,

four of them stressed that 2 heads and 3 heads are the two most likely outcomes of all. The following were four types of answers given by participants:

1. The number of heads is likely to be 2 or 3. (5/15: RY, QK, RM, PK, AQ)
2. The number of heads could be 0, 1, 2, 3, 4, or 5 but 2 and 3 are the most likely outcomes. (5/15: DE, WG, KU, HH, FX)
3. The number of heads could be 0, 1, 2, 3, 4, or 5. (3/15: PV, OZ, MP)
4. The theoretical number of heads is 2.5. (1/15: YW)

Participant GG responded "a lot of heads" and appeared to have difficulty comprehending the problem.

For question 2, all participants agreed on the indeterministic nature of the answer. 13 of them claimed that one would need to conduct experiments with the coin to answer the question. Five participants suggested tossing the coin many times and computing the relative frequency of heads. If this value was close to 0.5, then the coin could be fair. Three participants (PK, KU, HH) proposed something similar with an additional hypothesis test on the relative frequency of heads. Participant FX connected question 2 with question 1 and suggested that one repeatedly toss the coin five times to obtain a sampling distribution of the number of heads out of five tosses. If the distribution was similar to the theoretical binomial distribution for the number of heads when tossing a fair coin five times, the coin was probably fair. GG and MP mentioned tossing the coin many times but did not discuss how that would help determine the answer. Similarly, DE and AQ suggested conducting experiments but did not specify the experiments and how they would help address the question. In total, there were six basic responses:

1. Toss the coin many times and see if you get heads about half of the time. (5/15: PV, WG, QK, RM, OZ)
2. Collect a sample of coin tosses and perform a one-sample $z$-test on the proportion of heads. (3/15: PK, KU, HH)
3. Conduct experiments to collect empirical data and test fairness. (2/15: DE, AQ)
4. Toss the coin many times and see the results. (2/15: GG, MP)
5. We do not know. (2/15: RY, YW)
6. Toss the coin five times and repeat it many times. See if the distribution of the number of heads follows the theoretical distribution for question 1. (1/15: FX)

**Responses to Part 2**

The following were similarities that participants identified between the two questions. Some participants noted more than one of the relationships below:

1. Both questions are related to the fairness of a coin. (10/15: PV, DE, WG, RM, PK, KU, HH, FX, OZ, MP)
2. Both questions are about tossing a coin. (8/15: WG, RY, QK, GG, PK, KU, YW, MP)
3. Both questions need to use the theoretical probability of getting a head when tossing a fair coin. (5/15: PV, QK, KU, HH, MP)
4. Both questions have indeterministic answers. (2/15: KU, AQ)
5. Both questions use binomial distribution. (1/15: DE)
6. Both questions are based on independent coin tosses. (1/15: WG)

**Responses to Part 3**

The following were differences that participants identified between the two questions. Some participants noted more than one of the relationships below:

1. In question 1, fairness is the condition, but in question 2, fairness is the question. (6/15: DE, RM, KU, HH, FX, OZ)
2. We know the coin is fair in question 1, but we do not know if the coin is fair in question 2. (4/15: WG, RY, PK, YW)
3. Question 1 is more theoretical and computational, and question 2 is more experimental and conceptual. (3/15: PV, QK, GG)
4. Question 1 does not require data to answer, whereas question 2 requires data to answer. (1/15: DE)
5. Question 1 is about descriptive statistics, and question 2 relates to inferential statistics. (1/15: KU)
6. Question 1 is about the number of heads, and question 2 is about the fairness of a coin. (1/15: AQ)
7. In question 1, the expected number of heads is a long-term prediction of the results; in question 2, the expected number of heads is a reference value for testing the fairness of a coin. (1/15: MP)

**Responses to Part 4**

Except for participant YW who was unsure about the relationship, all other 14 participants identified probability as the connection between the concepts underlying these two questions. In particular, participants OZ and WG rationalized that question 1 used

theoretical probability to generate a list of possible numbers of heads where each answer came with a pre-determined theoretical probability. DE and FX referred to that probability as the binomial distribution for the number of heads when tossing a fair coin five times. Question 2 was more open-ended but could be answered using the experiment from question 1. One tosses the coin five times and records the number of heads. One repeats such an experiment a significant number of times and obtains another list of the numbers of heads where each answer is associated with a relative frequency count or the empirical probability. By applying the law of large numbers, which states that as the number of trials increases, the empirical probability shall approach the theoretical probability, one compares the two lists of possible numbers of heads and their corresponding probabilities. If the empirical probabilities are close enough to the theoretical ones, the examined coin is probably fair. To increase the objectiveness when making a decision, PK, KU, and HH recommended using a hypothesis test on the relative frequency of heads.

**Mathematical Thinking**

The first question is a mathematical question whose results can be predicted using a binomial distribution. For all possible numbers of heads out of five tosses, one can provide the exact probability or theoretical frequency count out of a fixed number of trials for each of them. For example, YW's response of 2.5 heads is purely a mathematical answer with no context. The most popular response of 2 or 3 heads is another example of a mathematical explanation. Even though these participants pointed out that getting 2.5 heads is not possible, they failed to see other less likely numbers of heads. In mathematics, one seeks the most

"accurate" answer to the problem. In this case, participants with mathematical thinking sought the most likely answers, the number of heads with the highest probability, to question 1.

**Statistical Thinking**

The second question is a statistical question that requires data collection. As 13 participants responded, one needs to conduct some sort of experiment with the coin in order to find the answer. They demonstrated the awareness of data collection for question 2. For nine participants who provided a complete plan for their experiment, six of their rationales were classified as statistical thinking. The remaining three, RM, QK, and PK, responded 2 or 3 heads to the first question. Their rationales were classified as mathematical thinking instead because they implemented a mathematical model without considering data production or context. Furthermore, when addressing part 4, probability was considered by 14 participants as a tool for statistics. By applying the law of large numbers, OZ, WG, and FX demonstrated their high level of statistical thinking. Their mathematical thinking in theoretical probability from question 1 is connected to their statistical thinking in empirical probability from question 2.

### Item 8: Statistical Problem

The seventh statistical problem tested participants' knowledge of confidence interval.

> Lindsey wants to use a confidence interval to estimate the difference in the proportion of females and males at her high school who have taken an honors class. She randomly selects 50 females and 50 males from her school and asks each one if he or she has taken an honors course. Of the 50 females, 23 responded yes. Of the 50 males, 19 responded yes. A 95 percent confidence interval for the difference in the proportion of females and males at her school who have taken an honors class is $0.08 \pm 0.19$.
>
> 1. Interpret the confidence interval in the context of this study.
> 2. The principal at Lindsey's school is interested in the results of her study but suggests that she increase the sample sizes to 100 females and 100 males.

> What effect will increasing the sample sizes have on Lindsey's confidence interval?

**Responses to Question 1**

There were three ways in which participants interpreted the confidence interval.

1. They interpreted the confidence interval numerically by describing the point estimate, the margin of error, and the endpoints of the interval. (4/15: QK, GG, AQ, YW)
2. They interpreted the confidence interval contextually by reasoning whether the observed difference in the proportion is statistically significant (whether 0 is included in the interval). (7/15: PV, DE, RY, PK, HH, FX, OZ)
3. They interpreted the confidence interval both numerically and contextually. (4/15: WG, RM, KU, MP)

However, not all participants provided the correct response. PV concluded that there was a statistically significant difference in the proportion based on the interval. GG's numerical interpretation of the confidence interval kept mistaking "mean" for the parameter to be estimated.

**Responses to Question 2**

All participants concluded that the confidence interval would become narrower:

1. As sample size increases, the margin of error will decrease because of the formula so that the confidence interval will become narrower. (6/15: PV, WG, GG, RM, PK, FX)
2. As sample size increases, the estimate's precision will increase so that the confidence interval will become narrower. (6/15: QK, KU, AQ, YW, OZ, MP)

3. As sample size increases, the margin of error will decrease so that the confidence interval will become narrower. (2/15: DE, RY)
4. As sample size increases, the sampling error will decrease, and so will the margin of error. Consequently, the confidence interval will become narrower. (1/15: HH)

All rationales were based on the change of the margin of error as the sample size increases. The first rationale explained this change using the mathematical formula of the margin of error. The second and fourth rationales interpreted the change using the statistical definition of the margin of error. The third rationale simply stated the change as the direct cause of a narrower confidence interval.

**Mathematical Thinking**

Numerical interpretation of the confidence interval is considered mathematical. Many participants were able to provide an answer that reported the confidence interval mathematically. For example,

> AQ: The confidence interval tells us that the estimated difference in the proportion of honor students in each gender is approximately 8% with a margin of error [of] 19%.
>
> KU: . . . the difference in the proportion of females and males at the school who have taken an honors class is between −0.11 and 0.27. Since it includes zero, there may not be any difference in the proportion.

Even though these statements are mathematically correct, they are statistically incorrect. These interpretations did not consider the possibility that the calculated confidence interval may not include the true difference at all. KU admitted that interpretations like his provided little evidence of students' conceptual understanding of confidence intervals. Anyone who knew how the math worked would probably be able to write results like these with minimal

knowledge of statistics. In fact, many of her students who managed to write mathematically correct sentences like AQ's failed to answer contextual questions that require an understanding of what a confidence interval means. They lacked an understanding, for example, of whether or not the confidence interval $[-0.11, 0.27]$ suggests a significant difference in the proportion of female and male honor students. Therefore, in KU's response to question 1, her second sentence addressed the practical implication of the confidence interval.

Additionally, in responses to question 2, six participants analyzed the change in the confidence interval's width mathematically. Based on the formula of the margin of error, the sample size is in the denominator of the fractional formula. As the sample size increases, the fraction decreases based on the mathematical property of fractions. This understanding reveals a solid mathematical knowledge about how the denominator's change affects the value of a fraction when its numerator is fixed.

**Statistical Thinking**

The first type of statistical thinking occurred in the interpretation of the confidence interval. In the responses to question 1, 11 participants provided the practical implication of the given confidence interval and demonstrated their statistical thinking. WG applied probability in his interpretation and stated that,

> There is a 95% probability that the interval from $0.08 - 0.19$, which equals $-0.11$, to $0.08 + 0.19$, [which ] equals 0.27, actually encompasses the true difference in proportions of male and female honor students.

Though WG demonstrated his probabilistic thinking, this interpretation incorrectly expressed the confidence level as the probability of the interval capturing the true parameter based on the current interval. The correct way to phrase WG's idea should be "There is a 95% probability

that the calculated confidence interval from a *future experiment* will actually encompass the true difference in proportions of male and females honor students."

Another common misstep related to this interpretation is making a statement about the probability of the parameter being captured by a given interval. For instance,

> GG: There is a 95% chance that the true "mean" lies in this interval.
>
> PV: I allow 5% of the time that the true proportion will not be in the interval, and it is still not unexpected.

These interpretations are incorrect because, under a frequentist model, one cannot assign a probability to the population parameter being estimated.

RM explained the confidence interval differently using the concept of repeated sampling:

> The confidence interval $0.08 \pm 0.19$ captures the true difference in the proportion of females and males at the school who have taken an honors class. And if this procedure is to be repeated on all possible random samples of the same size from the population, the proportion of all computed confidence intervals that will capture the true proportion difference would be approximately 95%.

The first part of this interpretation is incorrect because it is not guaranteed that the calculated interval based on the current sample includes the true difference. However, the second part of the interpretation about the repeated samples is correct. This idea of repeated sampling was discussed by almost all participants when they were asked explicitly what the 95% confidence level meant in a subsequent interview question. However, when interpreting confidence intervals, many participants did not specify the actual meaning of confidence level. Instead, they used words, such as "I'm 95% confident," "with 95% confidence," or "at 95% confidence level," that are generic and ambiguous.

The second type of statistical thinking related to the effect of sample size on the confidence interval. Seven participants described the effect of sample size on the precision of the estimate, indicating a statistical understanding of the margin of error. Particularly, HH interpreted the margin of error using its statistical definition–the maximum sampling error or the maximum difference between the sample estimate and population parameter.

### Item 9: Statistical Problem

The eighth statistical problem tested participants' knowledge of $p$-value.

> A farmer conducted an experiment to find out whether a new type of fertilizer would increase the size of tomatoes grown on his farm. The farmer randomly assigned 10 tomato plants to receive the new fertilizer and 10 tomato plants to receive the old fertilizer. All other growing conditions were the same for the 20 plants. At the end of the experiment, the mean weight of tomatoes grown with the new fertilizer was 0.4 ounce heavier than the mean weight of the tomatoes grown with the old fertilizer.
>
> 1. Describe one method that the farmer could have used to randomly assign the 20 plants into groups of 10 each.
> 2. Based on the results, the farmer is convinced that the new fertilizer produces heavier tomatoes on average. Briefly explain to the farmer why simply comparing the two means is not enough to provide convincing evidence that the new fertilizer produces heavier tomatoes.
> 3. To test whether the difference of 0.4 ounce is statistically significant, a statistician calculated a $p$-value of 0.31. Based on the p-value, is there convincing evidence that the new fertilizer produces heavier tomatoes than the old fertilizer on average? Explain.

**Responses to Question 1**

The most popular method among all participants was using a random number generator. The farmer first assigns a label for each plant. Then, the farmer randomly selects 10 plants from all labels and assigns them to receive the new fertilizer. This random selection could be made manually or with the help of technology:

> PK: I would name each plant, which could be numbered, and write them on index cards, put them in a black bag and shuffle them, blindly draw 10 cards from the bag, and assign these plants to the experiment group. The rest will be the control group.
>
> HH: In Excel, we can ID 20 plants with randomly generated numbers or names, then use its built-in "[data–]data analysis–sampling" to randomly select 10 plants to receive the new fertilizer. All unselected plants will receive the old fertilizer.

PK used the classic example of "drawing without looking from a black bag." Many other participants also picked this method. However, they did not mention "mixing" or "shuffling" before selection:

> KU: Give some cute names to the plants, like John, Jane, and so on. Write these names on index cards and put them in a bag. Blindfold and draw 10 names out of the bag. These lucky plants will receive the new fertilizer. The remaining plants will receive the old fertilizer.
>
> FX: Draw 20 cards from a deck of playing cards and link them to all plants in some way. Then place these cards in a hat and draw ten cards without looking. Those 10 plants linked to these cards will receive new fertilizer. The other 10 will receive old fertilizer.

Based on the grading rubric associated with this problem, KU and FX would only receive partial credit. However, as verbal responses, it was understandable that participants could have omitted "mixing" and "shuffling" unintentionally.

The following methods also involved a random number generator, but they did not specify how exactly the 10 plants would be randomly selected:

> PV: Write 20 numbers for the plants and randomly select 10 out of 20 to receive the new fertilizer.
>
> DE: We can randomly generate 20 numbers in Excel, then pick 10 randomly.
>
> YW: Label all plants and randomly draw 10 of them. They will receive the new fertilizer. The rest will receive the old fertilizer.

In addition to random number generator, some participants used a fair coin to determine the group assignment:

> RY: Place 20 plants in a line on the ground. Give each plant a number from 1 to 20 with no repeats. All the odd numbers will be one group, and all the even numbers will be the other group. Toss a fair coin. If it is a head, the group with odd numbers will receive the new fertilizer; and if it is a tail, the group with even numbers will receive the new fertilizer.
>
> OZ: For each plant, the farmer flips a fair coin to decide whether or not it receives the new fertilizer. If it's head, the plant receives the new fertilizer. If it's tail, then the old fertilizer. Repeat it until we have either ten heads or ten tails, then we stop.

However, each method has its own flaw. RY's method did not specify how each plant was assigned with a number, so randomness is questionable. In OZ's method, no matter which plant she began with, there would always be plants left not tested with the coin, and thus the process is not fair to all plants.

Additionally, two more methods relied on the physical location of the plant:

> GG: Don't look at it and choose the first 10 plants to put into the first group. The rest will go to the second group.
>
> MP: After planting all tomatoes, the farmers could give the new fertilizer to every other plant in his view and the old fertilizer to the rest.

These two methods are incomplete because the process of planting these 20 plants is not mentioned, yet whether or not these two methods are random depends on the decision of these plants' initial physical locations.

**Responses to Question 2**

Participants argued that 0.4 ounce heavier was not much, so it did not seem practically significant:

> WG: The mean difference here, 0.4 ounce, is really not a lot. It could easily occur by chance. I will ask the farmer to think about whether it makes sense to pay extra, well, depending on how much more the new fertilizer costs, for only 0.4 ounce increase in weight. Does it make sense practically?

Additionally, as a single mean difference, 0.4 ounce heavier was insufficient to describe the two fertilizers' comparative result. The farmer also needed to consider sampling variability:

> PV: The variance is unknown. Maybe the fertilizer is negatively affecting the plants. 0.4 ounce might seem better, but it's only the center, the average of all differences, some positive, some negative. If the variance is big, there could be many negative differences as well.

As a result, more experiments should be conducted:

> FX: . . . 0.4 is too close to 0. I will conduct another round of the same experiment.
>
> OZ: If the farmer starts a new batch and repeats the experiment, he may get a completely opposite result, such as 0.4 ounce lighter. I think what really matters is [to] see how large a difference will actually indicate that the new fertilizer is better than the old fertilizer.

In correspondence to OZ's response, some participants claimed that more statistics were required in order to determine the statistical significance of 0.4 ounce heavier:

> HH: The two means obtained are sample means which are the point estimates of the true population mean. Since they are estimates, they could be overestimating or underestimating. In other words, they are not exact. So we do not compare point estimates. We find the margin of errors for point estimates and compare the confidence intervals at a reasonable confidence level.
>
> AQ: (jokingly) The farmer needs to hire my statistics students and find $p$-value.

So it was not evident that the new fertilizer worked better.

**Responses to Question 3**

All participants responded that there was no convincing evidence that the new fertilizer worked better due to the $p$-value reported being greater than a common significance level, i.e. 0.05. In addition, participants DE and GG added "fail to reject the null hypothesis" in their response. Furthermore, five participants (WG, RY, RM, FX, MP) explicitly concluded that 0.4 ounce heavier was not statistically significant. Only participant PV provided some flexibility in her response:

> No, because the $p$-value is too high. It usually needs to be lower than 0.05 or 0.01. But it also depends on the field in which you are working. For example, in the medical field, even a $p$-value of 0.01 is not going to be acceptable. But in marketing, a $p$-value like 0.31 here may be acceptable.

**Mathematical Thinking**

First, participants' responses to question 1 were based on the mathematical probability of each plant being selected. Second, in question 2, HH and AQ claimed that the farmer needed to use confidence interval or $p$-value to justify the observed difference, which required an appropriate mathematical model. Third, in responses to question 3, all participants demonstrated mathematical thinking when comparing the $p$-value to a typical significance level. Everyone recalled the decision-making rules of $p$-value correctly.

**Statistical Thinking**

First, responses to question 1 suggested that most participants were aware of the process of random assignment in experiments. Second, in responses to question 2, many participants mentioned that the observed results could occur by chance alone, revealing that they were aware of variability in sampling and variability introduced by random assignment. HH demonstrated a higher level of such statistical thinking. She articulated the difference between comparing two means and two confidence intervals for the mean. She also stressed that when comparing sample means in statistics, one needed to take data production and sampling variability into account, which was different from merely comparing two averages in mathematics. Last, when addressing question 3, participant PV considered the context of the $p$-value and offered a response that was not strictly abode by textbook rules, demonstrating her original statistical thinking in understanding the importance of context in statistical analysis.

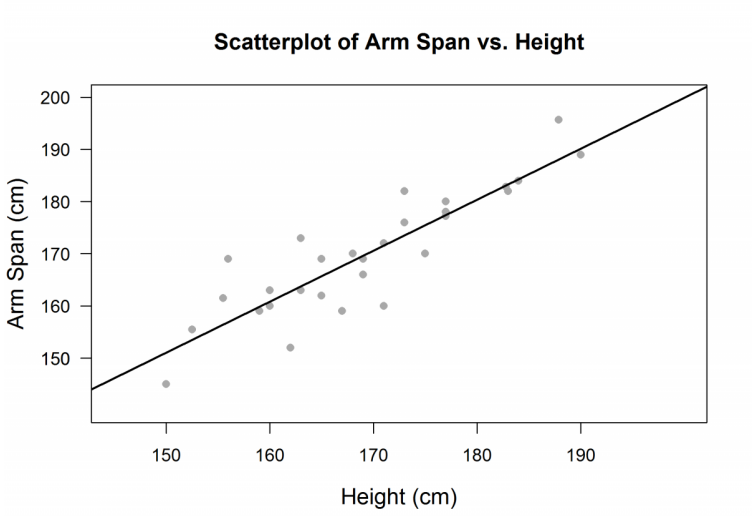
### Item 10: Statistical Problem

The last statistical problem tested participants' knowledge of simple linear regression.

The heights (in centimeters) and arm spans (in centimeters) of 31 students were measured. The association between $x$ (height) and $y$ (arm span) is shown in the scatterplot (see Figure C.4). The equation of the least-squares regression line for this association is given as following:

$$\hat{y} = 4.5 + 0.977x.$$

Figure 5.4. *Item 10: Height Versus Arm Span*




1. If Mike is 5 cm taller than George, what is the expected difference in their arm spans? Show your work.
2. Jane is 158 cm tall and has an arm span of 154 cm. Rhonda is 163 cm tall and has an arm span of 165 cm. Does the least-squares regression line give a more accurate predicted value for Jane or Rhonda? Explain.
3. Doug is 210 cm tall. Would you use this least-squares regression line to predict his arm span? Explain.

**Responses to Question 1**

At first glance, it may seem that the problem is missing key information. According to YW, "We need to know their heights to answer this question." But, many other participants

managed to conclude that their actual heights were irrelevant. One could find the answer via $0.977 \cdot 5$, but with different rationales. OZ and HH applied the definition of the slope in a linear regression equation which represents the rate of change:

> OZ: The slope of the regression line represents the rate of change. So the answer is 0.977 times 5 which is a little below 5.
>
> HH: For every 5 units increase in $x$, the height, the $y$, or the arm spans is predicted to increase by 0.977 multiplied by 5 which equals 4.885 centimeters.

KU and WG used the concept of shifting and scaling:

> KU: Because it's linear growth, the only part that determines the difference in their predicted heights is the multiplication part of the equation. So the answer is simply $0.977 \cdot 5 = 4.885$ centimeters.
>
> WG: The linear equation involves shifting and scaling. Shifting is plus 4.5; scaling is times 0.977. When plugging in Mike and George's heights, the difference in their arm spans will only be determined by the scaling factor and the difference in their heights, and that is $0.977 \cdot 5 = 4.885$ centimeters.

PK worked out the mathematics behind shifting and scaling:

> Assume George's height is $g$, then Mike's height is $g + 5$. Plug these two heights into the regression equation respectively and subtract:
>
> $$4.5 + .977(g + 5) - 4.5 - .977(g) = .977(5) = 4.885.$$
>
> Oh wow, so the only thing that matters is the slope. Ah, wait, that's the rate of change. That's why.

The second approach consisted of picking up heights for Mike and George and then plugging them into the regression equation to find the expected difference in arm spans:

> PV: Ok, so the equation tells us that, for every change in $x$, the change in $y$ [$\hat{y}$] is 0.977. Assume George's height is $x$, and Mike's height is $x + 5$. Then plug in and calculate $y$ [$\hat{y}$] for both George and Mike. To find the difference, you subtract.
>
> AQ: Assume George is 100 centimeters, then Mike is 105 centimeters. I can substitute their heights into the given equation to find out the expected arm span for each of them. Then subtract.

Despite the misuse of $y$ in PV's approach, her approach was essentially the same as PK's. The difference is that PV did not solve the problem but provided a plan only. As a result, she missed the chance to discover the math behind shifting and scaling. Similarly, AQ picked mathematically convenient heights and did not carry out his computation.

**Responses to Question 2**

The most popular approach used the regression equation. One finds their predicted arm spans which can be compared to their actual arm spans. The one with the less absolute difference has a more accurate prediction. Out of nine participants (PV, WG, RY, QK, RM, PK, AQ, FX, YW) who described this idea, only WG, RM, and PK did the math and concluded that the regression line gave Rhonda a more accurate predicted value. For instance, the following work was completed by RM:

$\hat{y}_{\text{jane}} = 4.5 + 0.977 \cdot 158 = 158.866$; residual $= |154 - 158.866| = 4.866$.
$\hat{y}_{\text{rhonda}} = 4.5 + 0.977 \cdot 163 = 163.751$; residual $= |165 - 163.751| = 1.249$.

The second approach used the scatterplot and the regression line. One plots the given information as two ordered pairs of coordinate points on the graph. The point that has the minimal vertical difference from the regression line has a more accurate prediction. A total of five participants (PV, DE, HH, OZ, MP) proposed this idea, but only HH and DE carried out the plan and solved the problem. They also concluded that Rhonda's predicted arm span was more accurate.

Participant KU provided a unique way of finding the solution:

> My guess is Rhonda. If we round the slope 0.997 to 1, then, based on the equation, we know that the predicted arm span should be 4.5 centimeters greater than the given height. Jane's arm span is 4 centimeters smaller than her height. Thus the equation won't give accurate predictions. In Rhonda's case, at least her arm span is greater than her height, and the difference is 2 centimeters, which is close to 4.5.

KU's approach essentially simplifies the math behind the first approach and allows a faster comparison result.

**Responses to Question 3**

Five participants (PV, RY, QK, GG, PK) responded yes and claimed that by plugging Doug's height into the regression equation, his predicted arm span could be easily computed. However, the other 10 participants responded "no." They argued that the approach would not be appropriate. For instance, HH reasoned that,

> Based on what I see from the scatterplot, the equation was created using heights between 150 and 190 centimeters. Doug's height is not within this range, so I will not use the line to predict his arm span.

In their opinion, the linear model was only valid for heights between 150 and 190 centimeters.

**Mathematical Thinking**

When addressing question 1, PK, PV, and AQ applied algebraic thinking. Since the problem asked them to predict arm spans (which depended on heights) but provided the difference in heights only, it is natural for them to use variables or convenient numbers to represent Mike's and George's heights. Among the three of them, PK demonstrated his high level of mathematical thinking in discovering the math behind shifting and scaling. PK had a brief "a-ha" moment while working on this question. At the end of the interview, PK credited his mathematical knowledge with allowing him to understand statistics more:

> Personally, I often find math useful for understanding certain things in statistics. I don't like to believe something [a] textbook tells me without seeing it from my perspective, so I like to do the math. That's how I learned when I was a student. If my students are able to work out the math behind it like I did, they won't need to memorize rules. And we know there are a lot of rules in statistics.

Additionally, KU and WG revealed their solid mathematical knowledge of shifting and scaling. They were able to pierce through the disguise of the problem and find the answer without setting up concrete heights for Mike and George. In question 2, most participants showed their mathematical thinking by using the regression equation. Particularly, KU applied mathematical techniques to simplify the computation. Lastly, participants who ignored the context of the regression equation in question 3 and plugged in 210 without questioning revealed their mathematical thinking by seeking a numerical answer when it is possible mathematically.

**Statistical Thinking**

In question 1, OZ and HH applied their conceptual understanding of the slope in the regression equation, demonstrating their statistical thinking about the rate of change and its context in statistics. In question 2, five participants used the visual representation of residuals to find the answer, revealing their statistical understanding of the least-squares regression line. When addressing question 3, 10 participants were aware of the validity of the linear model, demonstrating their statistical thinking about data production and context.

### Statistical Response Thinking Map

Table 5.5 demonstrated the overview of coded participants' thinking to nine statistical problems during the interviews (item 2 to item 10 from the interview handout). Statistical thinking, mathematical thinking, a mix of both, or neither were represented by orange, blue, olive green, and white, respectively. A darker orange or blue indicates strong evidence of statistical or mathematical thinking. The table was ordered by $ST$, a value that quantifies the tendency of a participant's primary thinking during the interview. A positive $ST$ indicated a tendency of statistical thinking, whereas a negative $ST$ implied mathematical thinking. A $ST$

Table 5.5. *Statistical Response Thinking Map*

| ID | $ST$ | 2 | 3 | 4 | | | 5 | 6 | 7 | | | | | 8 | | 9 | | | 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HH | 3.22 | | | | | | | | | | | | | | | | | | | | |
| OZ | 2.90 | | | | | | | | | | | | | | | | | | | | |
| DE | 2.53 | | | | | | | | | | | | | | | | | | | | |
| MP | 1.12 | | | | | | | | | | | | | | | | | | | | |
| KU | 0.87 | | | | | | | | | | | | | | | | | | | | |
| RM | 0.18 | | | | | | | | | | | | | | | | | | | | |
| PV | -0.18 | | | | | | | | | | | | | | | | | | | | |
| RY | -0.66 | | | | | | | | | | | | | | | | | | | | |
| FX | -0.75 | | | | | | | | | | | | | | | | | | | | |
| WG | -0.80 | | | | | | | | | | | | | | | | | | | | |
| AQ | -0.87 | | | | | | | | | | | | | | | | | | | | |
| QK | -2.01 | | | | | | | | | | | | | | | | | | | | |
| YW | -2.08 | | | | | | | | | | | | | | | | | | | | |
| GG | -2.65 | | | | | | | | | | | | | | | | | | | | |
| PK | -3.10 | | | | | | | | | | | | | | | | | | | | |

around 0 suggested either a mix of both types of thinking or a lack of either thinking. For a more detailed thinking map that included explanations of how $ST$ was computed as well as additional coded responses and rationales, see Appendix E.

### Hypothesized Aspects of Teachers' Thinking in Statistics

This section introduced a framework (Table 5.6) that distinguishes statistical knowledge from mathematical knowledge for teaching statistics and aims to categorize prominent aspects of mathematics teachers' thinking when teaching introductory statistics. The framework includes four columns. The first column lists the four components in a statistical problem-solving process (Franklin et al., 2007). The second and third columns state the identified mathematical thinking and statistical thinking in each component. The last column classifies the type of difference based on a review of the literature. One other colleague who had some experience in coding reviewed the framework and helped categorize examples of each type of thinking.

In the discussion which follows, examples taken from responses will be given for each type of thinking in the framework.

Table 5.6. *Hypothesized Aspects of Teachers' Thinking in Statistics*

| GAISE framework component | Mathematical Thinking | Statistical Thinking | Type of Difference |
|---|---|---|---|
| Formulate Questions | Formulate a (survey) question that doesn't require data collection and expects a deterministic answer. | Formulate a (statistical) question that requires data collection and evokes variability in response. | Problem solving process, variability, data production |
| | | Identify the correct population of interest based on a statistical question. | Context, data production |
| Collect Data | Collect data through a non-random sampling method that recognizes the additive view of a sample. | Collect data through a sampling method designed for recognizing the multiplicative view of a sample and reducing variability in data. | Problem-solving process, variability, data production, probabilistic thinking |
| Analyze Data | Plan and implement a mathematical model and attempt to solve the problem without consideration of data production or context. | Plan and implement a mathematical model and attempt to solve the problem with consideration of data production and/or context. | Context, data production, probabilistic thinking |
| | Analyze data and verify solution that adopts mathematical formulas and properties. | Analyze data and verify solution that accounts for context and variability in data. | Variability, context, transnumeration |
| Interpret Results | Interpret results about the sample by restating the original problem, reporting numerical results, or about the population by applying textbook definitions and rules. | Interpret results that make generalizations about the population with the inclusion of context and variability in data. | Variability, context, data production, transnumeration, probabilistic thinking |

**Formulate Questions**

Formulating question is a process unique to statistical problem-solving. While many problems in mathematics are given, one needs to formulate a question to start a statistical study. Also, there are particular types of questions that need to be used as the basis of a statistical study. First, one needs a statistical question that data can address. Second, to collect data, one may need to develop survey questions that will generate enough responses to answer the statistical question. Telling the difference between these two types of questions is the key to initiating a sound statistical study.

Table 5.7. *Examples of Teachers' Thinking in Formulating Questions*

| Mathematical Thinking | Statistical Thinking |
|---|---|
| Item 2: "How many students attend the school?" is a survey question. | Item 2: "How many text messages do students at the school send per week?" is a statistical question. |
| | Item 2: "How many hours does each class at the school meet per year?" studies a characteristic of the school, not of students at the school. |
| | Item 2: "Do students at this school have higher test scores than students in other schools in the district?" requires a random sample of students at the school and a random sample of students from other schools in the district. |

Despite the high number of participants who answered item 2 correctly, most of them revealed that they lacked deep knowledge of statistical questions and survey questions. It was

apparent that WG and GG, who chose both option B and option C, were not able to distinguish the different populations of interest from these two problems. In option B, the population was the school, and the characteristic under study was the number of students. The answer to this question was deterministic because the population only included one school. However, in option C, the population was all students at the school, and the characteristic under study was the number of text messages sent per week. Chances were, not all students would give the same answer. So it was very likely that there would be sampling variability in collected responses. YW and MP failed to recognize that the given condition, a sample of random students from the school, was insufficient to address option D because one needed a new random sample from other schools in the district. DE, who demonstrated a solid knowledge of this topic, articulated that option C, as a statistical question, required a random sample from all students at the school and expected variability in responses. In contrast, other options all studied an entirely different population.

**Collect Data**

Collecting data is another unique process in statistical problem-solving. In mathematics, a problem rarely requires the problem solver to collect data. However, in statistics, because most statistical questions address characteristics of a population that is typically estimated and inferred, one needs to collect a representative sample from the population, study the sample, and make inferences about the population. Furthermore, conclusions made based on a representative sample are more convincing than those made on biased samples. In data collection, the participating mathematics teachers' thinkings about samples of the population mainly splitted into two views.

Table 5.8. *Examples of Teachers' Thinking in Collection of Data*

| Mathematical Thinking | Statistical Thinking |
|---|---|
| Item 3: The student should take a random sample of students entering the library instead–the sample is a subset of the population. | Item 3: The students should take a random sample of students from all students, not just those entering the library–the sample mimics pertaining characteristics of the population. |

For example, in her rationale to item 3, YW selected option B, "the students should take a random sample of students entering the library instead," and demonstrated her additive view that a sample was part of a population, only. In contrast, RY selected option C and claimed that only option C included both students who entered the library and students who did not, which was a highly relevant characteristic of the population that needed to be included in the sample. Such a view is considered a multiplicative view of a sample in that a sample is "a mini version of the population." Though both option B and option C introduce a random sampling method that promises fairness mathematically to all subjects in their corresponding population, option C will more than likely produce a more representative sample for its inclusion of students who do not enter the library.

**Analyze Data**

Both mathematics and statistics involve the process of rigorously analyzing a problem using what's given. In statistics, that process usually requires the use of a mathematical model. The participating mathematics teachers revealed two different ways when implementing a mathematical model for a statistical problem.

Table 5.9. *Examples of Teachers' Thinking in Analyzing Data I*

| Mathematical Thinking | Statistical Thinking |
|---|---|
| Item 7: Use binomial distribution to address that the expected number of heads when tossing a fair coin 5 times is 2.5, 2, or 3. | Item 7: Use binomial distribution to address that the possible number of heads when tossing a fair coin 5 times are 0, 1, 2, 3, 4, 5. |
| | Item 7: Apply the law of large numbers to test the fairness of a coin. |

For example, in item 7, the first question asked the number of heads when tossing a fair coin five times. YW provided a purely mathematical answer of 2.5 heads which could be justified by the binomial formula $\mu = np$ where $\mu$ is the expected number of heads, $n$ represents the total number of tosses, and $p$ is the theoretical probability of getting a head when tossing a fair coin. YW implemented the correct mathematical model but failed to consider the context of the problem. Five mathematics teachers did recognize the context and concluded that the number of heads could not be 2.5, so they answered 2 or 3, which were the two most likely outcomes based on probability. However, they failed to mention less likely outcomes and ignored other possibilities in data production. The second question of item 7 asked whether a randomly picked coin was fair. OZ, WG, and FX not only adopted binomial distribution but also considered data production and the context of the problem. They described a coin toss experiment that applied the law of large numbers to connect theoretical probability from the first question to empirical probability from the second question, which asked whether a randomly picked coin was fair.

Additionally, mathematics teachers analyzed the data by applying different kinds of knowledge.

Table 5.10. *Examples of Teachers' Thinking in Analyzing Data II*

| Mathematical Thinking | Statistical Thinking |
|---|---|
| Item 4: Estimate center and variability of a dataset from a histogram. | Item 4 & 6: Recognize and understand the effect of extreme values on descriptive statistics. |
| Item 4: Use "average" instead of "mean." | Item 4: Use statistical terms for measures of center. |
| Item 5: Consistency in data is equivalent to consistency in frequency. | Item 5: Consistency in data is not equivalent to consistency in frequency. |
| Item 6: Connect standard deviation to mean absolute difference and coefficient of variation. | Item 6: Choose the appropriate descriptive statistics based on the shape of a given distribution. |
| Item 8: Use mathematical meaning of the margin of error to explain the effect of sample size on confidence interval. | Item 8: Use statistical meaning of the margin of error to illustrate the effect of sample size on confidence interval. |

Some mathematics teachers analyzed the problem by applying their knowledge of mathematical formulas and properties. For instance, in item 4, all of the mathematics teachers proved their ability to estimate measures of center and variation based on a histogram and compare two distributions by their estimates. What's more, 10 mathematics teachers used the word "average" to describe the center of a distribution without explicitly specifying which measures of the center they referred to. When responding to item 6, seven participants (DE, WG, RM, KU, HH, FX, OZ) used the mathematical formula for standard deviation to justify that standard deviation could be used to compare groups with different sample sizes. At a higher level, RY's and PK's justification connected standard deviation to other concepts in statistics (the mean absolute difference and coefficient of variation). Similarly, in item 8, six mathematics

teachers (PV, WG, GG, RM, PK, FX) applied the mathematical formula for the margin of error to predict the effect of increasing sample size on confidence interval. As PK stated, "I often find math useful for understanding certain things in statistics."

Some other mathematics teachers analyzed the problem by applying their knowledge of context and variability in data. Take item 5 as an example: when comparing the consistency of the responses between grade 7 and grade 8, many mathematics teachers mixed up the frequency with the actual data value and selected grade 7 incorrectly. To avoid the issue, KU and HH transformed the bar graph back to its original data form and confidently concluded that grade 8 had the most consistent choice, "aquarium." In item 8, HH argued that a larger sample size would decrease the maximum sampling error, which is also known as the margin of error, resulting in a narrower confidence interval. By connecting the width of the confidence interval, or margin of error, to the maximum likely sampling error, HH interpreted the concept of margin of error from a statistical point of view. Additionally, based on responses to item 4 and item 6, many mathematics teachers were able to illustrate the effect of outliers on descriptive statistics, especially the mean and median, and choose the appropriate measures of center and variation based on the shape of a given distribution.

**Interpret Results**

One important aspect when interpreting statistical results is the need to look beyond the data. In mathematics, the results are deterministic for a fixed set of conditions and assumptions. In statistics, results are computed based on a small group of subjects, estimating some property of a much larger group of subjects. From the microcosmic angle, these results are also deterministic for a fixed small group of subjects. But, in order to make inferences about

the larger group of subjects, which is the goal of statistical analysis, one needs to view statistical results from the macroscopic angle in which statistical results are indeterministic. There is more than one small group of subjects from the larger group. Each small group will produce possibly different estimates of the property of the larger group. If one looks beyond the data, one will soon realize that these results from small groups are not isolated. When enough of these results are collected, theoretically, estimates from all small groups form a distribution that describes all possible estimates based on small groups with distinct parameters such as central tendency and variability. One can use this theoretical distribution to evaluate the observed estimate from a single small group and determine if a statistically significant difference could exist between the estimate and the true value.

There are many examples in which mathematics teachers demonstrated a microcosmic view of statistical results. For example, in item 4, 11 mathematics teachers cited the conclusion from question 2 to address question 3, stating that since the mean of mile times for half-marathon runners was less than the mean of mile times for 5K runners, it was reasonable to conclude that the mile time of a person would be less when that person ran a half-marathon than when the person ran a 5K. They did not realize that the two dotplots they used to compare the mean were produced from two groups of different runners and that they cannot be used to approach a question that addresses the same group of people.

In item 8, some mathematics teachers did not include the confidence level when interpreting the confidence interval. A statement such as the one below is mathematical and deterministic. It does not express the possibility that the confidence interval may fail to capture the true difference in the proportion of male and female honor students:

Table 5.11. *Examples of Teachers' Thinking in Interpreting Results*

| Mathematical Thinking | Statistical Thinking |
|---|---|
| Item 4: Use measures of center to conclude question 3. | Item 4: Acknowledge how data were produced for each histogram; it is insufficient to use the current data to address question 3. |
| Item 8: Interpret confidence interval numerically by describing the point estimate, the margin of error and/or the endpoints of the interval. | Item 8: Interpret confidence interval contextually by reasoning whether the observed difference in the proportion is statistically significant. |
| Item 8: Interpret confidence level using a textbook definition. | Item 8: Interpret confidence level in context. |
| Item 9: Compare $p$-value to common significance level to identify statistically significant results. | Item 9: Recognize that 0.4 ounces heavier could occur by chance and there is sampling variability in the experiment; recognize that statistically significant results based on $p$-value are not absolute. |
| Item 10: Interpret regression equation algebraically. | Item 10: Interpret regression equation statistically. |

AQ: The confidence interval suggests that the estimated difference in the proportion of honor students in each gender is approximately 8% with a margin of error [of] 19%.

However, that does not mean that an interpretation of the confidence interval, such as the one below, that includes the numerical results of the interval, the confidence level, and its practical implication must be better.

RM: The confidence interval $0.08 \pm 0.19$ captures the true difference in the proportion of females and males at the school who have taken an honors class. If

> this procedure is to be repeated on all possible samples from the population, the proportion of all computed confidence intervals that will capture the true proportion difference would be approximately 95%. Since this interval includes 0, there may not be a difference at all between the proportions of the two genders.

One needs to be careful when making a deterministic statement about a calculated confidence interval as RM did above. Though RM's interpretation did reveal more evident statistical thinking when compared to AQ's.

Similarly, in item 9, though all mathematics teachers were able to apply the rules of $p$-values to conclude that 0.4 ounces heavier could occur by chance, only PV considered the context of $p$-values and discussed the possibility of breaking textbook rules on some occasions. Unfortunately, the initial intention of $p$-value, given by R. A. Fisher, which recommends that a low $p$-value is a sign that the experiment should be repeated, was not articulated by any participants.

Finally, mathematics teachers demonstrated different types of interpretation of the regression equation in item 10. Some mathematics teachers treated the equation as another algebraic linear equation in the slope-intercept form and applied algebraic thinking in solving question 1 and question 2. In contrast, other mathematics teachers applied the statistical definitions of the slope and $y$-intercept to seek answers differently. Even though both ways of thinking can lead to the correct solution, they indicate different levels of statistical understanding of simple linear regression. Such a difference in the understanding was further revealed in question 3 where all incorrect responses (PV, RY, QK, GG, PK) came from mathematics teachers who solved the first two parts algebraically.

A complete overview of all mathematics teachers' coded responses to statistical problems from the interview was included in Appendix D and Appendix E.

# Chapter 6

# RESULTS: RESEARCH QUESTION 2

In this chapter, the second research question was addressed: *With a general knowledge of various types of statistical technology options, how does mathematics teachers' statistical thinking or mathematical thinking affect their way of teaching?* In each section, mathematics teachers' responses to pedagogical questions associated with each corresponding statistical problem (item 2 to item 10 from the interview handout) were reported and analyzed based on the specialized knowledge from the SKT framework I (Groth, 2007). Regarding how thinking affects teaching, examples of mathematics teachers' pedagogical choices were matched with their thinking based on research question 1. Effects were summarized in three areas: topic coverage in statistics, delivery methods in the class, and student assessments.

## Item 2: Pedagogical Questions

The pedagogical questions relating to item 2 focused on the concepts of statistical questions and survey questions.

> Q1: When you teach introductory statistics, do you mention the concept of statistical questions? How do you introduce the concept of statistical questions? For instance, do you tell students the difference between statistical questions and survey questions? If so, how do you explain the difference?
>
> Q2: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

### Responses to Question 1

Out of 15 participants, 10 claimed that they had never mentioned or taught the concept of statistical questions. Among the rest, participant GG misunderstood the question and

provided an irrelevant response. Based on the remaining four participants who did mention statistical questions in their teaching, there were two primary ways in which they used to introduce the concept and differentiated it from the concept of survey questions:

1. A statistical question requires a large group of people to answer it, whereas a survey question only requires one individual's answer. (3/5: QK, KU, HH)
2. A statistical question requires data collection, whereas a survey question does not. (1/5: OZ)

The following was a short conversation between QK and the researcher while discussing pedagogical question 1 for item 2:

> QK: I often tell my students to pay attention to the subject of the question. The subject in a statistical question is usually a large group of people, while the subject in a survey question is usually "you" or the respondent. So for a statistical question, you are expecting to get a list of various answers. But for a survey question, you are expecting to get just one answer from each respondent. I usually assign a semester-long group project to my students that tests knowledge that students learned from the entire semester. And the first part is about choosing an appropriate statistical question. My students will see a list of questions to choose from. Every group must choose one from the list. I remind them to think about how they will collect data before making a selection.
>
> Researcher: Where do you get your statistical questions?
>
> QK: I get my questions from the instructor's resources that come with different textbooks. I have taught with different textbooks and explored what's available there. Some are pretty good. They have these project ideas that have detailed descriptions of what students need to do, almost step-by-step. Sometimes I change questions to those I found on the internet. I do ask my students to write their own survey questions because they need to collect data, which is mostly a survey, in-person or online.
>
> Researcher: Do you evaluate their survey questions? Or help students improve them?
>
> QK: If they ask yes, but no.

QK was one of the few participants who answered item 2 correctly and provided a rationale that involved variability in statistics. His good understanding of statistical questions and

survey questions allowed him to administer group projects that mimicked a real-life statistical analysis. Unfortunately, his students did not get the chance to formulate statistical questions, and their survey questions were not assessed formally. Near the end of the discussion, QK told the researcher that he had given more open-ended group projects when he first started teaching. However, it didn't end well since most students lacked the ability to create a sound and study-worthy statistical question, and so did he. With a Bachelor's degree in math and an EdD degree in math education, QK only took two statistics courses. He did not recall learning about statistical questions. Therefore, he deemed that he was not sufficiently prepared to teach the concept.

**Responses to Question 2**

While most participants had no clue how technology integration could be used to facilitate the teaching of statistical questions, participant DE, who answered the statistical problem correctly and demonstrated statistical thinking in his rationales, stated that technology could be used to both simulate the sampling process and generate data for a proposed statistical question, so students could continue using the same example for future topics in statistics. His actual implementation was discussed in the next section. Participant HH, who introduced the concept of statistical questions to her students, pointed out that a project based on option C would be challenging for students to implement since it was almost unrealistic to obtain a theoretically perfect random sample from all students at the school. She provided her own solution. In her class, HH asked her students to download a large dataset from the Census at School that could be used to answer a given statistical question. Then

students chose a random sampling method to draw a random sample from the dataset for further analysis.

### Item 3: Pedagogical Questions

The pedagogical questions for item 3 focused on the concepts of random sampling.

> Q1: When you teach introductory statistics, how do you explain the concept of random sampling to students?
>
> Q2: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

When introducing the concept of randomness, the most popular response was the textbook definition that a random sample is formed from a population where every member has an equal chance of being selected. Many participants claimed that they had used textbook examples to discuss different random sampling methods. It was typically done in class discussion or group work. Students investigated different ways a sample could be collected and discussed whether or not they sounded fair. Some participants provided their own examples in real-life context:

> RM: I propose three plans to give students extra credit and ask whether they sound fair. Here are my plans. No. 1, give extra credit to students whose last name starts with M; No. 2, give extra credit to students who scored 100 on exams; and No. 3, give extra credit to students randomly selected by a random number generator. Usually, students would say plan 2 and plan 3 are fair. If I am lucky, maybe one student whose last name starts with M would shout out that the first plan is the fairest of all. Then I tell my students that change of plans–plan 2 now gives extra credit to students who scored 60 and below on exams instead. Many students will realize that plan 2 is not fair anymore. But the truth is plan 2 is never fair as long as I determine the required score. We then conclude that fairness in statistics means all students need to be given an equal chance to receive extra credit.
>
> RY: I use a coin to illustrate the idea of random–that you cannot predict what the next toss result is. You may get it right, and you may get it wrong. And there is

> really no pattern in guessing. Sometimes one or two students got really lucky in guessing, and I had to keep tossing until they got it wrong. Those were fun moments of teaching.
>
> AQ: I put everyone's name in a bag and draw five names with replacement multiple times to demonstrate that each selection could be different; some people get selected multiple times, some don't at all, but everyone has a chance to be selected.

All three activities aimed to illustrate the concept of randomness but with different aspects. RM's example focused on the equal theoretical probability of every member being chosen from the population or the mathematical aspect behind random sampling. Her example also revealed how the word "fair" is typically interpreted differently in statistics from its everyday use. RY and AQ revealed the uncertainty in random occurrence or the statistical aspect behind random sampling. Additionally, RY deemphasized the textbook definition of a random sample and conveyed to her students the challenge of obtaining a perfectly random sample.

**Responses to Question 2**

Quite a few participants talked about integrating educational technology, such as web applets or software, to demonstrate random sampling through simulation of the actual sampling process:

> RM: I have seen some Pearson web applet or online program that demonstrates different sampling techniques and asks students to determine whether the technique is random. They use text, voice-over, and animation to verify or correct students' choices, which is nice.
>
> HH: In Excel, I randomly generate one integer from 1 to 10 and repeat 1000 times. I plot a histogram of the frequency distribution or relative frequency distribution of all results and ask my students to tell me what they see.
>
> QK: Some applets can simulate random sampling, such as coin flip, dice toss, random number digit generating. Students can easily study the distribution.
>
> AQ: I know that you can use Google Sheets to randomly select students using the random number generator.

> RY: Originally, no idea. But after exploring the applets you showed, I feel like a simulation, like tossing a coin or drawing a card, will be great for demonstrating concepts like random. Students can explore by themselves, and it could be a productive teaching episode.
>
> PK: Could be. I saw many sampling applets on the RC website [Rossman & Chance Web Applets].

It is worth noting that the technology evaluation part of the study helped some participants to see potential in educational technology for teaching statistics. Among all those who were in favor of technology integration in random sampling, participant DE discussed in detail how he had used web applets from StatKey to introduce the concept of the sampling distribution for a mean:

> I frequently use applets from StatKey to let my students observe how randomness affects sampling variability and how population determines sampling distribution for a parameter. For instance, there is an applet where I can click a button to generate one sample and find its sample mean. The applet automatically plots the mean as a dotplot. I can do it multiple times to show how different samples result in different sample means–students may find each sample independent of another and independent of the population with unpredictable sample characteristics. I can also easily click another button to generate 1000 samples instantly and visually plot the sample means for all of them on the same graph to display the sampling distribution for a mean. Some students who see the connection between the sampling distribution for a mean and the population may change their mind and conclude that what [they] previously considered as unpredictable now actually follows a particular pattern.

DE commented that many modern textbooks in statistics now promoted the use of technology to facilitate students' conceptual understanding of difficult topics such as sampling distribution. A good understanding of sampling distribution would serve as a precursor for subsequent topics in inferential statistics. He was delighted to see that not only one but many technology options were recommended in the textbook, with step-by-step tutorials available to both instructors and students.

### Item 4: Pedagogical Questions

The pedagogical questions for item 4 focused on teaching histograms.

> Q1: If you are showing these two histograms along with the context to your students, what statistical concepts would you address in this example? How would you teach these concepts using this example?
>
> Q2: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

For question 1, the following are the topics indicated by participants:

1. Descriptive statistics from a histogram (15/15)
2. Outliers and their effects on descriptive statistics (13/15)
3. Histogram (10/15)
4. Different distributions (9/15)
5. Comparison of histograms (7/15)
6. Histogram versus bar graph (2/15)
7. Applying context to interpret descriptive statistics from a histogram (1/15)
8. Questioning data production before comparing histograms (1/15)

The first five topics are standard ones in many introductory statistics classes. In particular, based on the participants' frequent mention of outliers and their effects on descriptive statistics, topic 2 appears to be popular among modern statistics instructors. The delivery of topic 2 can be either mathematical or statistical. The upcoming section on item 6 looks at the subject from a pedagogical perspective.

Topic 6, histograms and bar graphs, is usually a common area of misunderstanding among students. But only participants PV and WG expressed interest in teaching this topic. PV

stated that many of her students were in need of formal instruction on these two types of graphs so that they could create and interpret each of them. WG commented that his departmental curriculum for statistics placed a heavy emphasis on quantitative analysis while gradually getting rid of qualitative analysis tools such as bar graphs. He still taught bar graphs, but it never became more sophisticated than reading a bar graph created from a one-way table.

Topics 7 and 8, contexts of descriptive statistics of histograms and questioning data before creating histograms, were mentioned by two different participants, DE and HH, respectively. Both of them answered question 3 correctly, demonstrating their high level of statistical thinking. According to DE, numerical computation of descriptive statistics is not a skill because many modern technologies can complete that instantly. When students leave school and perform statistical analysis in real-life, no one will compute everything by hand. However, being able to interpret numerical results based on conceptual understanding is a skill, in DE's opinion. Likewise, HH told the researcher that she could not wait to give item 4 to her students and see how many students would fall into the trap in question 3. In her opinion, tasks like item 4 that do not ask students to perform heavy calculations but rather test students' interpretation of descriptive statistics are problems worth assigning to introductory statistics students.

**Responses to Question 2**

Regardless of participants' background in educational technology, all responded "yes." The major benefit of technology in this topic was the convenient construction of histograms from data. Technological aid allowed both instructors and students to focus on the

interpretation of the graph. Additionally, participant DE added that certain technology enabled a dynamic view of histograms that benefited students even more:

> Absolutely. By using technology, for example Excel, one can easily create a histogram from a dataset. Technology also allows the user to change the number of bars or adjust the class width. Another benefit of technology is that one can easily add or remove data points and see the visual change in the existing histogram without recreating the graph. I think this creates an open learning environment for statistics learners. In the past, I used Excel to illustrate the effect of outliers on the mean and median. By adding or removing the outlier, students will see both the visual change in the histogram as well as the numerical change in the calculated mean and median. I find out that my instructions are much more effective when students see statistical ideas unfolding in front of their eyes.

### Item 5: Pedagogical Questions

The pedagogical questions for item 5 focused on consistency in categorical data.

> Q1: In your teaching, do you ever address the comparison of consistency in categorical data to students? If so, how?
>
> Q2: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

Three participants recalled that they had addressed the comparison of consistency in categorial data:

> DE: In my class, I give examples that create clustered bar plots and ask students to use the graph to answer questions. They are usually free-response questions, kinda like the runners' problem [item 4]. But I don't use the word consistency. Instead, I use the word "same" or "different". I tell students to look at the height of the bars.
>
> YW: I use visual representations such as histograms to illustrate.
>
> KU: I used to, maybe 10 years ago, when I was using a different textbook. Students were asked to construct a two-way table and create a bar graph based on that. I believe we were examining whether or not the two categories in the two-way table were associated. I think the new textbook that I'm using emphasizes more on histograms and quantitative data analysis.

DE used words that made more sense to his students when introducing the concept of consistency. However, it is clear that DE did not understand item 5 correctly, and he used heights of bars to justify his incorrect selection. As a result, he applied the same incorrect strategy to teach consistency in categorical data to his students. YW failed to recognize the difference between bar graphs and histograms. She might be referring to her teaching experience of comparing the consistency of numerical data. Similar to KU, many participants who had taught statistics for years expressed their concern that current instruction in introductory statistics primarily focused on quantitative analysis. There might be a shortage of introduction in the qualitative analysis at the entry-level:

> RM: Not really. But as far as I'm concerned, bar graphs are the most common graphs that students will plot in the future. Based on my teaching experience, however, many textbooks nowadays are deemphasizing the importance of bar graphs. For instance, the suggested syllabus provided by one of my colleges where I teach doesn't even include chapters on contingency tables. On the exam, bar graphs are rarely tested. But my students will definitely see histograms on the exam. And you know what happens next–many of my students see bar graphs and call them histograms.

**Responses to Question 2**

Out of 15 participants, six participants responded "maybe," claiming that technology could assist in visualizing categorical data more easily. But because they had not done it or addressed the comparison of consistency in categorical data this way, they were not sure how effective it would be. Among the remaining nine participants, seven responded "yes," and supported the view that technology could efficiently transform data, either raw or tabular form, into a bar graph:

> QK: Technology can help create bar graphs faster so students can focus on interpreting the graph instead of graphing by hand.

> OZ: I use Excel to create bar graphs. I do like to use technology to visualize data. It's more efficient.

Two participants provided examples of how to use technology to help students see that grade 8 was the correct answer to the question in item 5:

> RM: For instance, I sometimes survey my students at the beginning of the class. I collect responses from students online, usually [a] Google Form. Let's say I have some categorical responses to two questions, where do you want to visit, the aquarium or the zoo? Do you like statistics? Yes or no only. I will copy all responses into Excel. I create two pivot tables there with the responses and frequency counts. I ask students to look at the tables and find out which question has the most consistent responses. To check their answers, I will create bar graphs for both questions and show [them] to students. In the end, students get to investigate the consistency in categorical data from both the table and graph. I think if the data from this problem are put in a table, maybe more students will pick grade 8 correctly.
>
> HH: It is easier to explain this problem if I show students the original dataset which is just a list of "zoo"s and "aquarium"s. They will see that Grade 8 has the most consistent answer because it has the most "aquarium"s on the list. If I present the data to students and throw the same question, I believe most students will answer grade 8. Then I will create the bar graph that you see here using technology and blow my students' minds! That sounds fun!

Both RM and HH answered item 5 correctly and demonstrated their high level of statistical thinking in data production and transnumeration. While most other participants gave a general statement that technology would help visualize the data, RM and HH narrated a more vivid story about how technology integration could actually facilitate their teaching.

### Item 6: Pedagogical Questions

The pedagogical questions for item 6 focused on the instruction about outliers.

> Q1: In your teaching, do you ever address the effects of outliers on descriptive statistics? If so, what do you address and how?
>
> Q2: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

All 15 participants responded "yes." There were five ways in which participants addressed the effects of outliers on descriptive statistics:

1. Use fake dataset with and without outliers to illustrate:

   > WG: I wrote down a few numbers such as 1, 2, 3, 4, 5 on the board. Then, I guided students to compute descriptive statistics such as the mean, median, range, IQR, and maybe standard deviation–It's a good opportunity to review what they have learned. Once that's complete, I add an extreme number, like 1000, to the list and ask students to repeat all calculations. Once they see the before-and-after descriptive statistics, my students often manage to conclude the effects of outliers on their own.

   PK used similar approach. QK used a web applet to demonstrate that with the addition of extreme values to an arbitrary list of numbers, the numerical change is greater in mean than in median. While this may be a standard method to illustrate this topic, DE told the researcher that using fake data in a statistics class is not a good practice since it "rips context off numbers and sucks the life out of statistics."

2. Use textbook examples to illustrate:

   > WG: The textbook has some nice-looking graphs that visually rationalize and summarize the effects of outliers on the mean and median. I like to present them to my students when they have completed group work [on generating descriptive statistics from fake data].
   >
   > KU: I address the effects with textbook examples to show that, with the inclusion of outliers, some descriptive statistics, like the mean and standard deviation, change more than others, like the median and IQR. I tell my students that not all descriptive statistics are meaningful. It depends on the shape of the distribution.
   >
   > AQ: The textbook that I use has real-life examples of how outliers affect the measure of centers and which one should be appropriate to use in each scenario.

KU commented that she preferred textbook examples because they were closely related to students' homework problems, and her students thanked her for discussing "meaningful" examples in class. Similarly, FX and GG also used textbook examples to illustrate, but GG primarily focused on how to identify outliers using mathematical formulas.

3. Use real-life examples to illustrate:

> PV: I give links to news reports to show students why, for certain things such as household income or house sales prices, they use the median instead of the mean to describe a typical value. It makes it easier for my students to understand that when we include super-rich people or their house sales prices, the average or the mean of all income or sales prices will no longer be representative of everyone's.
>
> RM: I like to talk about their Blackboard grades. On Blackboard, you can set to display the average and the median for any graded item. Typically, the mean is much lower than the median. I ask my students why, and some of them tell me that there could be zeros in the calculation which lowers the average. Since it's related to grades, I can grab the attention of most of my students. Students will learn to compare their grades to the median instead of the average, which is a good way for them to remember and understand the effect of outliers on the mean and median as well as its practical implications.

Unlike KU, RM believed that her students found some textbook examples hardly relatable. She claimed that her students were typically interested in any discussion related to their grades, so she tried to design the entire statistics curriculum based on students' grades.

4. Use visual representation to demonstrate:

> DE: I use boxplots. Students will see several boxplots with the mean labeled on each graph. Some boxplots have outliers, and some don't. Then, I will ask students to tell me how the distribution, which is determined by the boxplots, affects the positions of the mean and median in each boxplot. We end the discussion with conclusions for different situations.
>
> RY: I usually prepare a histogram to illustrate the effects of outliers on measures of center. My students are divided into groups. Each group will

> receive different kinds of histograms–normal, right-skewed, left-skewed, uniform, and other asymmetrical shapes. I ask each group to mark where they think the mean and median should be for each histogram. Then each group will present their work and explain why they chose to place the mean and median in such a way.

Both DE and RY adopted a student-centered teaching method that concentrates on students' open exploration and self-discovery in the topic.

5. Use visual representation with the aid of technology to demonstrate:

> HH: I know a website that does the job! It allows you to choose the initial distribution of points or even create your own, so it is possible that no students will face the same problem. The web page nicely displays the numerical and graphical comparisons of the mean and median under the given distribution. I usually let my students play with it before jumping to the final conclusion. Students need to see what they are learning. In my opinion, it greatly helps with their conceptual understanding.

Similar to DE and RY, HH let her students learn from exploring different distributions. By using the web applet, her students freely explored the effects of outliers individually or as a group. Thanks to the applet's customization feature, her students were able to investigate the concept with unlimited cases of outliers.

Even though it appears that most participants focused on the effects on the mean and median only, the effects on the standard deviation and IQR could be explored similarly in most cases. DE commented that since the standard deviation and IQR were more difficult to compute and visualize, one should implement technology if possible.

**Responses to Question 2**

Nine participants responded "yes" or "maybe." Among them, there were two primary uses of technology mentioned. First, one used technology to demonstrate the effects of outliers visually. It could be a web applet that displayed a distribution and computed descriptive

statistics as HH described above. Participant PK, who had not used technology for this topic, found an applet while exploring technology options mentioned in this study:

> PK: I used an applet from one of your sites. I forgot which one it was, but it's about descriptive statistics. I was able to generate random samples by clicking a button. And I had the option to display both the mean and median on a histogram or a dotplot. It's very easy, at least for me, to see the effect of extreme values on these two different measures.

Second, one used technology to demonstrate the effects of outliers numerically. Both RM and MP mentioned that Excel or Google Sheets computed descriptive statistics of a list of numbers instantly. By adding or removing a few extreme values, students could simultaneously observe the corresponding changes caused by these extreme values.

However, not all experience with technology was pleasant. Being a self-claimed Excel lover, RM shared a frustrating teaching moment in Excel when she taught the calculation of sample standard deviation by hand. First, she instructed her students to construct a calculation table in Excel, which breaks down the formula of standard deviation into separate columns. RM expected that the use of Excel would minimize the calculation burden on students, and the scaffolding of the table would help students understand the mathematical formula for standard deviation. However, as students started to fill in the table in Excel, RM noticed unexpected issues. First, many students were unaware of the auto-fill function in Excel, where one could easily repeat the same computation by dragging or double-clicking the lower right corner of the cell. This issue impeded students' use of technology to calculate the sample standard deviation efficiently. Second, for those students who knew this function, many of them had difficulty figuring out how to lock the cell where the mean was located, which was necessary to enable auto-fill in this type of calculation. As an unfortunate consequence, RM had to spend

extra time explaining the specific usage of Excel. Her students were perplexed by the incorrect numbers displayed in the table, which hampered their understanding of the formula. RM's experience led to pedagogical fidelity in which her real teaching experience mismatched with her initial pedagogical objective and expectation.

### Item 7: Pedagogical Questions

The pedagogical questions for item 7 focused on the instruction of probability in an introductory statistics course.

> Q1: Some statistics instructors choose to trim down the instruction of probability in introductory statistics courses. They claim that many probability topics are not related to the subsequent instruction of inferential statistics. What do you think?
>
> Q2: What role do you think probability plays in introductory statistics?
>
> Q3: How much probability do you teach to your students and why?
>
> Q4: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

There were different opinions when it comes to trimming down the instruction of probability. Some participants left out certain topics in probability but for different reasons. For example, In her teaching, RY sometimes skipped counting rules to be able to cover other topics required by the syllabus. She claimed that counting rules were rarely used in subsequent chapters. The only counting rule that she did talk about in detail was the combination formula which her students needed for the binomial distribution. PK and AQ stated something similar:

> PK: I remember reading something like that in the instructor's version of the textbook. The writer of the book recommended that instructors should skip certain sections in probability. Most of my students are not well prepared for statistics. Many of them can't even do basic arithmetic. So the course pace is slow

> sometimes, and I had to skip or speed up some sections so that I can cover everything required by the department.
>
> AQ: Though after teaching statistics for so many years, I do tend to trim down the instruction of probability, especially the theoretical part. One, my students are not math majors. They are not interested in the "why" part. Two, the newer edition of the textbook is trimming down and simplifying the coverage of probability topics for instructors.

It is interesting to see that RY, PK, and AQ all demonstrated mathematical thinking based on their response to question 1. And they all appeared to deem that excluding topics in probability was acceptable. DE also omitted topics in probability. In fact, he left out almost all topics in probability. DE explained that, in order to be enrolled in his statistics class, students needed to pass a course that taught students all the topics they needed in probability for his class. He would rather spend more time on new topics that were difficult for most students. As RM said,

> If probability has been taught systematically in a lower-level class or another math course before, I'd say it's okay to trim down the instruction of probability. I do see the need to do this, especially if there are lots of chapters to cover.

However, DE considered trimming down topics in probability dangerous since unprepared students would have a hard time learning many other topics in statistics. DE was not alone in concerning about excluding topics in probability. HH stated that,

> You can skip some topics but not all. For example, the $p$-value is essentially a probability. If students report a $p$-value that's not between 0 and 1, they either don't understand the $p$-value or the probability. But my bet is probably both. So students need to know the basics of probability. As far as I am concerned, not all topics in probability need to be taught. For instance, my department doesn't require us to teach Bayes' theorem. I think that can be skipped if time is an issue.

In addition to HH, many other participants, such as QK, KU, OZ, and MP, mentioned the close connection between probability and $p$-values. Because of this connection, these participants disagreed with trimming down topics in probability. WG expanded this connection to other topics in statistics:

> My two cents is that some ideas in probability, such as the rare event rule, or the law of large numbers, do reappear in later chapters of introductory statistics. Like the decision-making in hypothesis testing can be understood by the rare event rule. The graph of a normal distribution is essentially a result of the law of large numbers. When the instructor makes connections like these, students' conceptual understanding of statistics could be substantially improved.

There was also a practical reason why skipping topics in probability was not feasible for YW:

> I do spend a lot of time on probability because most of my students are struggling with probability. Unfortunately, there are many probability questions on the exam which is not made by me. We have standardized department exams. If I skip too many topics in probability, my students won't pass the exam.

In general, participants who discerned probability playing an important role in understanding inferential statistics and demonstrating statistical thinking tended to considered it unwise to teach fewer topics in probability.

**Responses to Question 2**

Almost all participants used the word "important" when talking about the role of probability in statistics. GG stated that probability introduced students to uncertainty in statistics. AQ declared that probability provided students with tools to justify sampling method in statistical analysis. YW mentioned that probability was widely used in decision-making in statistics. Particularly, many participants, such as HH, DE, RY, MP, KU, QK, explicitly stated that probability facilitated students' understanding of $p$-values or hypothesis testing.

**Responses to Question 3**

The following were the topics in probability that the participants said they taught:

1. Probability basics (15/15: PV, DE, WG, RY, QK, GG, RM, PK, KU, HH, AQ, RX, YW, OZ, MP)
2. Conditional probability (11/15: PV, WG, RY, QK, GG, KU, HH, AQ, FX, YW, OZ)

3. Probability rules (10/15: WG, RY, QK, GG, HH, AQ, FX, YW, OZ, MP)
4. Bayes' theorem (8/15: PV, WG, QK, GG, KU, AQ, YW, OZ)
5. Counting rules (6/15: GG, KU, AQ, FX, YW, OZ)
6. Rare event rule (4/15: QK, GG, RM, KU)

FX asserted that probability should be placed as a separate course taught before statistics to provide the necessary theoretical foundation for statistics. But PK held the opposite opinion and suggested that only the basics of probability needed to be covered.

**Responses to Question 4**

Many participants expressed approval with regard to using technology to teach probability. Mainly, technology was used to visualize a probability problem, so students were able to understand probability beyond text:

> PK: I remember there is an online program for the Monty Hall problem from one of your websites. Now I can use it to demonstrate the problem and ask my students to play as many times as they want. The program even records results so my students can see whether it is better to switch or not. And I believe that Tinkerplot[s] has some fun activities related to probability as well. Technology makes probability problems alive to students and easier for them to grasp.
>
> DE: For example, when calculating probability from a normal distribution, many of my students had trouble remembering when they need to subtract from one. So I introduced an online applet that shades the desired area under a normal curve. It not only shows students how geometry is connected to probability, but also helps students check their answers. For example, if the shaded area is more than half of the total area under the curve, their answer must be more than 0.5.

Additionally, technology was used to simulate experiments that are hard to accomplish in real-life, so students were able to observe and analyze long-term behaviors:

> QK: I use technology to simulate drawing a card or rolling a die. Like roll a pair of dice 100 times and find the sum. Technology is perfect for it. Students calculate the theoretical probability first and then observe the simulation to compare what they have predicted and what they have observed.

> RM: Like one time, I found this online program that generates random results of flipping a coin. But first, you need to enter the odds. Let's say the probability of getting a head is 25%. By observing the simulation, I asked my students to determine whether or not the coin seems fair. The program generates results in two forms, a table, and a histogram. I used this example to illustrate the rare event rule–if what we observed about the number of heads is far away from what we expected, our previous assumption that the coin is fair is most likely not correct. Thus the coin is most likely not fair. Later, I revisited the same example when teaching hypothesis tests.

Some participants recalled their experience exploring different technology options for this study and remembered seeing applets or programs investigating topics in probability. The most frequently mentioned technology option was Tinkerplots. However, with the knowledge of what technology could do for teaching statistics, WG still preferred using physical manipulatives for teaching probability:

> Personally, I haven't done any [technology integration]. But I did regularly toss dice and flip coins in my class. I'm an old school guy. When I toss a die, I'd like to physically feel the die going up in the air, see it slowly dropping, and hear it hitting the desk. Imagine you are watching it in slow-mo. And my students have loved tossing dice in class too. They interact with each other, not with a computer screen. I am aware that technology can toss a die thousands of times within a second, but what's the fun of that?

Finally, it's worth mentioning that KU, who had never used any technology to teach statistics, talked about her teaching experience with technology during the 2020 pandemic:

> You know, I am not very good at technology, so I'm, maybe, reluctant to use any. I don't want to call the tech support all the time and waste my students' time. See, I just recently got a smartphone, and I am still learning how to use it. However, with this COVID-19 pandemic going on, I have forced myself to learn new technologies that I never knew existed, like this Zoom thing. Technology is a great invention by humans. When it's used wisely, many great things can happen. But, to answer your question, I'm afraid I won't be able to give you anything useful.

The COVID-19 pandemic of 2020 forced many teachers and educators to transform their course delivery to a fully online or hybrid mode. That had pushed many statistics teachers to

implement technology in their teaching. KU was struggling in many aspects of online teaching when the interview took place in early 2020, but she did not give up on trying and learning.

### Item 8: Pedagogical Questions

The pedagogical questions for item 8 focused on the concept of confidence level.

> Q1: Some students have difficulty understanding confidence level. In your teaching, how do you approach this difficult topic? For instance, how do you tell your students what 95% confidence level means?
>
> Q2: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

Many participants used the textbook definition to address the confidence level:

> WG: I used an explanation provided by the textbook in which 20 confidence intervals have been constructed and plotted vertically on the same graph. The graph also shows that the true population parameter passes through 19 of all 20 intervals, that 95% of all confidence intervals constructed are accurate in encompassing the true population parameter.
>
> MP: The textbook describes a simulation that demonstrates what 95% means. In this simulation, we construct confidence intervals for all possible samples of the same size from the population. If we graph all confidence intervals together and mark the true population parameter with a straight line, we should see that the line goes through about 95% of all confidence intervals. The textbook gives a graph of one such simulation, which I showed to my students.

MP commented that students usually had difficulty understanding the simulation described by the textbook, but he did not know how to write a program to perform the simulation. Some other participants with experience in technology demonstrated the simulation described by MP. According to them, none of the technology options required its user to write a program:

> DE: I apply StatKey and use bootstrap CI [confidence interval] without first telling students the formulas because bootstrap CI will help students get an idea of what CI means. It also demonstrates the 95% CI graphically. 95% is about the interval. If

> you have 100 CI's, 95% means 95 of them will cover the true parameter. The parameter is unchanged, but the CI might vary depending on the sample.
>
> RM: I show my students an online program that performs a simulation of a particular number of confidence intervals based on a population. The program will then compute and illustrate the percent of all intervals generated that actually contain the true population parameter. If I initially set the confidence level to 95%, as the number of generated intervals increases, the percentage will get closer and closer to 95% and fluctuate at some point.
>
> HH: I use this website. It displays the percentage of intervals that contain the value estimated. I like how you have the options to customize population parameters and sample sizes.

Participant KU tackled the concept of confidence level from a different angle:

> I ask my students to estimate my age by giving me an interval. I tell them that the interval can be of any width. Most students usually report back with a width of 5 years. There are usually a few who would give me extreme wide or narrow intervals. If there is none, I would say one as an example. Then I select three intervals, one very wide, one normal, and one very narrow, and ask students that, out of these three interval estimates, which interval do you feel the most confident about estimating my age? I use this activity to illustrate how improving confidence level also increases the interval's width and provides less useful estimates.

KU's approach provided a more intuitive understanding of how confidence level affects the estimating process and results. However, this method was based on a misconception that confidence level represents how confident one feels about an interval estimate being accurate.

**Responses to Question 2**

In addition to using technology to explain what confidence level means, as discussed above, PV and RM added that technology could be used to investigate the effect of sample size on the width of a confidence interval. RM described a class activity in which she used Excel to assist her teaching of confidence intervals:

> A lot of my students are visual learners. They usually don't have a sense of what the interval means. Sometimes they write intervals where the lower bound and upper bound are reversed. Technology can visualize intervals for them and is

> extremely useful when you have multiple confidence intervals. For instance, in my confidence interval lecture, I ask my students to estimate the mean number of M&M candies, let's say the number of green M&Ms in each packet. I divide my class into groups and give them some samples of M&M candies. Sometimes, the lecture occurs during the Halloween season, which can't be more perfect. As the entire class, we ask all groups to merge their group samples into a class sample. By following the formula, my students construct a confidence interval for their group sample and another confidence interval for the class sample. This is basically the instruction part of confidence intervals in my lecture. Then, in Excel, I compute confidence intervals for all groups and the entire class. I project both numerical results and visual intervals to my class. They check their work and discuss what happens to the width of the interval when the sample size increases in the class sample.

RM further commented that this activity of hers might have violated many statistical assumptions and conditions for the proper implementation of confidence intervals. However, she considered it as a good practice for introductory statistics courses in which theoretical emphasis is usually less critical.

### Item 9: Pedagogical Questions

The pedagogical questions for item 9 focused on the concept of $p$-values.

> Q1: Some students have difficulty understanding $p$-values. In your teaching, how do you approach this difficult topic? For instance, do you ever address the conceptual understanding of $p$-values? If so, what do you address and how?
>
> Q2: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

Participants reported that they mainly addressed three aspects of $p$-values. First, they brought up the rejection rules of $p$-values in tests of significance. All participants mentioned the topic, but some chose to focus on the application of $p$-values only for different reasons. Participant RY was very honest about her lack of education in statistics and topics such as $p$-values:

> I didn't really explain what a $p$-value means because, to be honest, I have trouble understanding it too. I have never taken any statistics courses. The department assigned me to teach statistics when I first got hired as an adjunct. When I first taught statistics, I basically read the textbook and learned along with my students. From a mathematical point of view, the $p$-value is pretty simple. When the $p$-value is small enough, like less than 0.05, I teach students that we observe statistical significance and reject the null hypothesis. Otherwise, we fail to reject the null hypothesis. And honestly, that's all that the test asks students about anyways.

Similarly, QK and YW admitted that their students only needed to know the rejection rules of $p$-values for exams. FX stated that, sometimes, it was a learning choice made by students:

> To tell you the truth, I tried explaining [$p$-values] in the past, but it was not helpful. No student cared. My students prefer memorizing rules on how to use $p$-values to answer questions. Also, my instruction of $p$-values is usually near the end of the semester when many students' only concern is about what the exam will be like. The exam is usually non-conceptual, so very few of my students would actually pay attention to the conceptual part of my teaching. To help my students on their exams, I can't help but cut anything conceptual in my lecture. I know it's not ideal in statistical learning, but it's effective in statistical grades. My students are not STEM majors. They just want to pass the course and graduate. And I know how hard their life can be, especially this year [COVID-19].

FX commented that, throughout her twenty years of teaching experience at this college, she had continuously lowered her coursework difficulty to meet the new department standards and the new generation of students. With all the distresses happening amid 2020 and all the distractions students now have, she had learned to adapt and teach statistics in a way that her employers like, in a way that her students want, and in a way that she would probably disagree with twenty years ago.

Second, participants explained the definition of $p$-values. Regardless of the depth of the instruction, most participants gave a textbook definition of $p$-values similar to the one stated by the American Statistical Association (Wasserstein & Lazar, 2016):

> A $p$-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value. (p. 131)

However, PK stated that his students found the definition of $p$-values difficult to understand. To facilitate students' understanding, some participants provided alternative definitions of $p$-values:

> DE: If the default knowledge is correct, how likely your particular data will support that assumption.
>
> WG: Given that a defendant claims to be innocent, the $p$-value is how likely that the defendant is telling the truth. If the $p$-value is small, we reject the defendant's claim and conclude that the defendant is probably guilty. If the $p$-value is large, we fail to reject the defendant's claim and conclude that there is not enough evidence against defendant's claim. But, it doesn't mean that the defendant is innocent.

However, DE warned that using an alternative definition of $p$-values may lead to misconceptions about $p$-values as it could be misinterpreted. For example, PV used the same defendant definition and asked her students, "At what $p$-value do you want that person to go to prison? The one with [a $p$-value of] 1% or the one with [a $p$-value of] 5%?" It is dangerous to ask questions like these because students may use the $p$-value to measure and compare the size of an effect or the statistical significance of a result, which $p$-values cannot do.

There were two other alternative definitions of $p$-values reported by PV and KU that are worth discussing separately:

> PV: The $p$-value is the probability of accepting a hypothesis knowing that you are gonna be wrong that percent of time and being OK with it.
>
> KU: I use type I error to explain what a $p$-value measures. I start by explaining that $\alpha$ is the probability of incorrectly rejecting $H_0$. Let's say that $\alpha$ = 0.05. Assuming $H_0$ is actually true, the hypothesis test is repeated on different random samples of data from the same population; then we would expect $H_0$ to be incorrectly rejected 5% of the time. If $p$-value is less than $\alpha$, that means $H_0$ will be incorrectly rejected less than 5% of the time. In other words, $H_0$ will be correctly rejected more than 95% of the time. That suggests that the assumption $H_0$ is more than likely

> supposed to be correctly rejected. So we reject $H_0$ and conclude that $H_0$ is probably not correct.

These definitions may seem more straightforward to comprehend when compared to the textbook definition of $p$-values. However, both definitions incorrectly interpret the $p$-value as the probability of making a mistake when rejecting the null hypothesis, which is another common misconception about the $p$-value. The calculation of the $p$-value is based on the assumption that the null hypothesis is true, so its value does not measure the likelihood of either the null hypothesis or the alternative hypothesis being true or false.

In addition, GG and AQ used the rare event rule to facilitate their teaching of $p$-values:

> GG: I will give students a different example to explain the $p$-value. Like tossing a coin, what is the probability of getting a head? 0.5. Do it again, get another head. Now, the probability of getting two successive heads is 0.5 times 0.5; do it again, get another head; keep tossing it. Is it possible to keep getting heads? Now you may get suspicious that if you keep getting heads, the coin may not be a fair coin. So if you get something that's weird, I mean far away from your expectation, you are going to reject your assumption. The $p$-value is the probability of observing something extreme that you do not expect to observe when we assume something is true.
>
> AQ: I tell my students that the $p$-value is a probability, and it measures the probability of observing something equivalent to or more extreme than what is present in the sample when the null hypothesis or assumption is true, for example, drawing cards from a deck of cards. Assuming it's a fair deck of cards with an equal chance of getting red and black cards, if we keep getting black cards, since its theoretical probability is very low, we can conclude that our assumption is probably incorrect and that the deck is not fair.

The rare event rule states that, under a particular assumption, if we have observed that something implausible occurred, then the assumption is probably not valid. It is correct for GG and AQ to connect the rare events rule to tests of significance. However, there is still a difference between what the rare event rule measures and what the $p$-value measures. When applying the rare event rule, the probability of what's being observed is measured. However,

when defining the $p$-value, it not only measures the probability of what's being observed from the sample, it also measures the probability of those events that are more extreme than what's being observed. Therefore, extra caution must be taken when explaining the $p$-value using the rare event rule.

Third, participants reported giving their students a visual representation of the $p$-value. To further foster students' understanding of what the $p$-value measures, some participants visually depict the $p$-value and connect small $p$-values to empirical rules of the normal distribution:

> OZ: Students who learn about $p$-values are already familiar with the concept of significantly high or significantly low values using the normal curve–they usually fall on the two tails of a normal curve. Visually, a smaller $p$-value fills up one or two tails of a normal curve, indicating something significant.
>
> RM: I tell my students that the $p$-value is the probability of obtaining an effect that is equivalent or more extreme than the one observed in the sample data, assuming the truth of the null hypothesis. This definition is typically very abstract to my students. So I use an online program to generate a sampling distribution of, let's say, sample means of a dataset via bootstrapping. By marking a select sample mean on the histogram, the program shades the two-tailed areas representing regions equivalent to or more extreme than the given mean and computes the area, namely the $p$-value. By the empirical rule of a normal distribution, students will get a better picture of whether the observed result is within the "unlikely" region. If yes, it means the assumption or null hypothesis is probably not correct, and the observed result is statistically significant; if no, the observed result is not statistically significant, and it could occur by chance.

RM commented that whenever she solved a $p$-value problem in class, she always visually indicated the shaded area representing the $p$-value. Her students were required to provide the visual $p$-value in their work and check it against their computed numerical $p$-value. According to RM, this had helped her students validate their work before handing it in for grading.

**Responses to Question 2**

Most participants considered technology as a faster way to find $p$-values:

> RY: We can use technology to find $p$-values much faster. Also, we can teach students how to read statistical results from technology outputs, which, in my opinion, is probably the most practical use of $p$-values for them. I think being able to use technology to conduct statistical analysis is a plus when looking for jobs nowadays.
>
> QK: With technology, the $p$-value of different scenarios can be computed rather quickly so that students can focus on the interpretation of the results.
>
> PV: The $p$-value should be seen in real data and used to get conclusions in real studies. Technology can help with calculation, and use examples from real-life . . . [to look at] . . . how students interpret $p$-values [given] in journal articles.

Corresponding to PV's statistical thinking in context, she expressed her preference for teaching $p$-values using real-life data and interpreting $p$-values from journal articles. RY demonstrated her awareness of educational technology in statistics and its increasingly important role in the current job market. Even though she still taught statistics in a traditional way with minimum use of technology at the time of the interview, RY expressed her eager to learn new technology options and teach more up-to-date skills to her students.

Additionally, technology was appraised to visualize and quantify $p$-values:

> DE: On StatKey, there is a hypothesis testing menu for sampling distribution. By simulating repeated sampling and computing desired sample statistics, a $p$-value can be perceived as one- or two-tailed areas that are as extreme or more than the observed sample statistics from the first sample. This nicely introduces the $p$-value before going too formal.
>
> HH: This applet visualizes $p$-values, so students get a better picture of what they are computing. Students can change the hypotheses from a two-tailed to the left- or right-tailed and instantly see the change of the $p$-value both numerically and visually. Even though the applet computes the $p$-value for students, I always verify it in Excel with my students.

Based on their statistical responses to item 9, both DE and HH demonstrated a solid statistical understanding in the definition of the $p$-value and its connection to sampling distribution. As a result, they were able to apply technology simulations to introduce $p$-values differently.

### Item 10: Pedagogical Questions

The pedagogical questions for item 10 focused on the concept of simple linear regression.

> Q1: Some students find the hand computation of the linear correlation coefficient $r$ and regression line tedious and impractical in real-life (since many applets or spreadsheets can complete the calculation within seconds). What's your opinion on calculating $r$ and regression line by hand in a modern introductory statistics class?
>
> Q2: Is the question above similar to what you will normally assign to your students? If yes, could you give an example of a similar problem you have assigned? If no, could you give an example of a typical problem you would assign on this topic and illustrate the differences?
>
> Q3: Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Responses to Question 1**

More than half of the participants taught hand calculation of simple linear regression. Five participants (RY, QK, RM, YW, FX) stated that they had to teach how to calculate by hand because it was required on the exam made by the department. Though it might be true for other participants as well, DE and GG mentioned that important formulas in statistics were provided to students on the exam. Additionally, some participants pointed out that hand computation in statistics could help boost the confidence in students with low self-esteem:

> RY: In my class, it is necessary to know how to compute by hand because it's on the department exam. It's part of being in a math class. Students are expected to carry out some complex calculations. Plus, students will appreciate technology more when they have experienced hand calculation. When students finally obtain

> their answer, and it is correct, they gain a stronger sense of achievement and become more confident in the course. [Researcher: what about those who didn't get the correct answer or got lost in hand calculation?] I usually help them on-site and guide them through to the end. I want all my students to taste victory in math.
>
> KU: As far as I'm concerned, every student needs to experience the complexity of statistical computation by hand. Maybe hand computation will become obsolete tomorrow, but the experience of conquering something challenging and obtaining the final correct answer is something that low self-esteem students desperately need, especially in math classes. If I make something look too easy, students won't work hard on mastering it because they think they understand it all; if I make something look too difficult, students will give up too quickly because they don't believe they have it in them. I think that things such as formula computation of $r$ are something that most students can manage and follow through. To some students, it gives them an idea of doing the actual math in a math class successfully, which is something that technology can not provide.

In particular, AQ discussed his opinion on procedural instruction in statistics:

> I would say yes to hand computation. Some instructors think it's too procedural. But I don't think procedural is automatically bad instruction. It can lead to positive results too. Almost all of my students come from non-STEM backgrounds. They typically have bad experiences when taking math classes. I know this because I usually conduct a survey in our first class. One of the survey questions asks students to talk about their previous experience in math, anonymously, of course. Then I ask the same question again at the end of the semester. I want to see if their experience has changed after taking my class. Of course, you will have students who still had bad experiences. But, surprisingly, some students told me that they enjoyed computing the standard deviation by hand, or they loved the feeling after writing an entire page for hypothesis testing. So what I want to say is that, as educators, we should begin by teaching our students something that they are capable of doing, like following a recipe step-by-step. Once they see that they can accomplish something, they will be more inclined to learn maybe a bit more conceptual stuff.

A few participants claimed that computing $r$ promoted students' conceptual understanding:

> PV: I don't mind letting my students compute by hand. It promotes their number sense. It also allows students to see the dependency between different quantities. So I don't oppose it.
>
> WG: I did hand computation of $r$ for many years in my class. The breakdown of the formula is helpful for understanding the theoretical aspects of $r$. I believe it is important for students to apply the formula and understand the formula, especially if they are math majors or statistics majors.

> PK: Sure, technology is convenient, but it hides all the mathematical steps designed to lead to the final result. To my knowledge, most technology implementations involve a sequence of clicking buttons and entering functions. And I am not talking about those that require programming. So it's like cheating in a video game that you loaded someone else's saved progress and go straight to the ending. I'm not saying that students don't need to know how to do it with technology. They will still need to learn at some point. By learning how to calculate by hand now, students will be able to do statistical analysis with and without technology in the future. It's just more options for them to solve a problem.

On the contrary, MP thought the formula for $r$ was tedious with no pedagogical value:

> Many of my students are not very good with numbers, even with a calculator. They often get points taken off on the exam due to computation errors. But that has nothing to do with statistical knowledge. That's the ability to follow a formula and calculate. We all know the formula for $r$, what it looks like, especially if it uses the original data points. There is a simplified version using $z$ scores, but it is still tedious and, in my opinion, has no real value in enhancing students' understanding of the topic. So I strongly recommend it be simplified. I would suggest providing the value of $r$ to students and letting students use that to find the equation by hand. I think that's more reasonable. Or, like this problem, give students the equation right away.

OZ taught hand computation in class but recommended her students to use technology for statistics outside the class. Some participants stated that, in real life, it was improbable that students would conduct statistical analysis purely by hand. By using technology, students focused on the interpretation of results and addressed more conceptual questions in statistics, which was the real objective of learning statistics:

> RM: One college where I teach requires students to compute by hand on exams. I find it utterly meaningless. We should teach students how to use free tools that they have access to, like Excel or online programs that you showed us in the survey, to perform simple linear regression and obtain results. It's more meaningful to ask students questions based on the interpretation of the regression line. I think this problem is a very good example of what statistical exams should be like on this topic.
>
> HH: In Excel, instead of using built-in functions, I guide my students through the manual computation of $r$. I find that many students can interpret $r$ correctly but cannot find the correct value due to its complicated computation. What's worse is

> that the exam, which I do not make, usually asks students to find $r$ by the formula. When I grade, I grade based on students' ability to interpret, not their ability to compute. If they are going to conduct linear regression in the future, I think they will use Excel or other software to find $r$ and the regression line. It's the interpretation that they need to know to report the results. And we should really teach them how to interpret statistical results and address contextual word problems. What I'm about to say may sound cringey, but I believe that each number in statistics has a story. Our job, as instructors, is to unfold these stories to our students so that one day they can do the same on their own.

Based on the responses of the 15 participants, it appears that participants who demonstrated strong mathematical thinking (PK, PV, AQ, KU, WG) all supported or felt comfortable with teaching hand computation of $r$. In contrast, participants with prominent statistical thinking (OZ, HH) recommended calculating $r$ with technology.

**Responses to Question 2**

No participants reported that they had assigned similar problems to their students. Most participants assigned questions that asked students to compute $r$ and the equation by hand. Those participants primarily consisted of those who supported or felt comfortable with teaching hand computation of $r$. DE and GG assigned questions that asked students to find $r$ and the equation by reading from statistical outputs. DE stated that, in his opinion, students' ability to read statistical outputs was more important than their ability to compute statistics numerically. Instead of calculating $r$, RM assigned questions that asked students to estimate $r$. Additionally, many participants (DE, GG, FX, OZ, MP) have also asked students to interpret the equation's slope in words using the context of the problem.

Moreover, many participants reported that they assigned similar problems that had students predict the dependent variable using the regression equation. However, FX stated that

her prediction questions did not involve predicting more than one subject or the evaluation of residuals. DE mentioned that his prediction problem does not cover out-of-domain values.

**Responses to Question 3**

For some instructors, technology has changed their way of teaching:

> WG: In the old days, I collected data from my students and created scatterplots on the board by hand. It was fun but time-consuming. Nowadays, I use Excel to create most scatterplots. I like how easy it becomes to explore different relationships in class with my students.
>
> MP: Instead of showing pictures of scatterplots from the textbook, now I use technology to generate scatterplots randomly. I ask students to make predictions of $r$ or sketch the regression line on the screen. Students can click a button to self-check their predictions and move on to the next scatterplot.
>
> RY: I used to rely on textbooks or publisher's test bank to assign homework problems on linear regression. Sometimes my students told me that the answer keys were incorrect. But now, after learning how to do it in Excel, I can check answers using technology or create my own problems and answers.

Most participants agreed that the use of technology saved time on computation, enabled more examples to be discussed, and created more teaching ideas in class. In addition to the popular choice of Excel, participants also mentioned other technology options that they used for teaching linear regression, such as R (DE), Google Sheets (AQ), and a graphing calculator (FX). For the most part, technology fostered a learning environment that welcomed exploration and self-discovery among students:

> RM: Technology helps students find $r$ and the regression line much faster. Students can add more data points and observe how the position of different points affects the value of $r$ as well as the graph of the regression line.
>
> HH: Linear regression is a fun topic to teach. Before diving into the standard topics, like the complex computation of $r$ and regression equation by hand, I ask my students to explore linear regression here. It introduces many important concepts such as the regression line, how it is determined, the residual and how it is defined, the $r$, and what it measures or how it is affected by the scatterplot. I can

> even extend the topic to include a residual plot and show how it indicates that linear regression may not be a good fit for the points.

Out of the 15 participants, 14 managed to provide at least one example in which technology could be used to facilitate the teaching of simple linear regression. It seems that technology integration in linear regression was common among participating mathematics teachers. But a few mentioned that they learned only by exploring different technology options during this study:

> PK: I haven't integrated any in my teaching because I didn't know how. But now I remember seeing a few, what do you call them, . . . applets on websites that you gave us. They are related to linear correlation. I remember that you choose your data set from a drop-down menu and create a scatterplot instantly. It also displays other useful statistics like $r$, the equation, or the standard error.
>
> FX: I think it is in Tinkerplots that you can literally drag a data point to a different location on the graph and observe the instantaneous change in the regression line. That's simply amazing!

Participants with mathematical thinking tended to stress the convenience and efficiency of technology in producing results. Exploration-based learning appears to be a common pedagogical possibility for both participants with mathematical thinking and statistical thinking.

### Effects of Teachers' Thinking on Teaching

In this section, examples of mathematics teachers' pedagogical choices were given based on the hypothesized aspects of teachers' thinking in statistics (Table 5.6) from research question 1. In general, mathematics teachers' thinking affected their teaching in three areas: topic coverage in statistics, delivery methods in the class, and student assessment.

**Topic Coverage**

Item 1 on the interview handout asked participants to provide a list of topics that they normally covered in their introductory statistics course. Even though all participants taught

Table 6.1. *Topics Covered in Introductory Statistics Course*

| Topics | #Response | Percent |
|---|---|---|
| Sampling and data | 15 | 100.0% |
| Descriptive statistics | 15 | 100.0% |
| Probability topics | 14 | 93.3% |
| Discrete random variables | 15 | 100.0% |
| Continuous random variables | 15 | 100.0% |
| The normal distribution | 15 | 100.0% |
| The central limit theorem | 15 | 100.0% |
| Confidence intervals (one sample) | 15 | 100.0% |
| Hypothesis testing (one sample) | 14 | 93.3% |
| Inference from two samples | 7 | 46.7% |
| Linear correlation and regression | 15 | 100.0% |
| Chi-square test | 4 | 26.7% |
| The F distribution | 1 | 6.7% |
| Analysis of variance | 2 | 13.3% |

at the same college, they were adjunct instructors teaching in different locations. As a result, not all participants gave the same results. Based on the responses, there was a lack of coverage of later topics, which had to do with inferential statistics. It should be noted that this list did not reveal the subtopics that each participant covered. Two participants could choose the same

topic but cover different subtopics. Additionally, as some participants commented during the interview, some topics were embedded within other topics as a single chapter. For example, DE stated that the chi-square test was part of his instruction in hypothesis testing. Lastly, this list did not suggest the order in which these topics was taught by participants. For instance, HH taught linear correlation and regression before probability topics.

In addition to introducing all the required topics on the departmental syllabus, most mathematics teachers covered topics that they believed they understood at a deeper level. For topics that they were not familiar with, depending on whether or not the topic would be tested on the exam, they either chose to exclude them or cover them at a level that they felt they could handle. According to responses to the pedagogical questions concerning item 2, the concept of statistical questions was not popular among mathematics teachers. Four mathematics teachers (QK, KU, HH, OZ) who did teach statistical questions were able to tell the difference between statistical questions and survey questions, demonstrating their statistical understanding of the topic. While much of their instruction on this topic merely focused on the textbook definitions and examples, one mathematics teacher, QK, had mentioned assigning group projects that require students to create survey questions. However, students' survey questions were not formally assessed, nor did the students need to develop statistical questions. QK expressed his concern that his knowledge of statistical questions was not sufficient enough to guide his students through a complete statistical study that started with the formulation of a sound statistical question that also interested his students. To not confuse or mislead his students, QK chose to implement group projects made by textbook publishers and had statistical questions already written for students. QK attempted to

contextualize this topic, but most other mathematics teachers did not give it too much attention, mainly because they did not know the concept of statistical questions well enough.

The teachers with strong mathematical thinking tended to cover more mathematical topics. For instance, mathematics teacher PK, who demonstrated his solid mathematical knowledge in solving many statistical problems in the interview, preferred teaching the effects of outliers mathematically using numerical examples such as made-up numbers in item 4. When he was asked to justify comparing the standard deviation of groups with different sample sizes in item 6, PK brought up the concept of coefficient of variation, which normalized standard deviation when comparing groups with very distinct means. It was typical for PK to present the complete underlying math work to his students if he considered it necessary for his students. In item 10, PK also supported the importance of hand computation in statistics. In his opinion, technology neglected necessary mathematical steps that led to final results. Yet, those steps were part of the course and essential for anyone learning statistics, especially at an introductory level.

Similarly, the teachers with strong statistical thinking were inclined to cover more statistical topics. For example, when being asked what topics they would cover related to histograms like the one in item 4, DE and HH proposed topics 7 and 8 ("Applying context to interpret descriptive statistics from a histogram" and "Questioning data production before comparing histograms"). These two topics were different from the rest for their distinct statistical implications. DE and HH, who demonstrated a high level of statistical thinking in answering question 3, chose to teach students more conceptual topics related to histograms. DE considered the ability to interpret descriptive statistics from a histogram or technology output a rudimentary skill that all modern statistics students should obtain when they had

completed the course. Inspired by question 3 of item 4, HH was excited about giving the problem to her students. She believed it would teach her students an important lesson on paying attention to data production in statistics.

Probability was an interesting topic, and the teachers held very different opinions about it. RY, PK, and AQ, who demonstrated mathematical thinking in item 7, along with HH and OZ, who demonstrated statistical thinking in item 7, did not oppose trimming down topics in probability. YW, who gave the most mathematical answer of 2.5 heads to question 1, stated that she never skipped topics in probability and spent a substantial amount of time preparing her students for probability questions on the exam. QK, who also provided a mathematical answer of 2 or 3 heads, considered probability to be close related to inferential statistics and should not be trimmed down. However, mathematics teachers who demonstrated statistical thinking in item 7 all agreed that, at minimum, fundamental topics in probability should always be taught because many statistical concepts, especially in inferential statistics, depended on an understanding of basic probability.

One important application of probability in statistics is the $p$-value. However, based on 15 mathematics teachers' responses to item 9, the $p$-value was not an easy topic for mathematics instructors to teach in statistics, which agrees with findings from the existing literature (Franklin et al., 2015; Motulsky, 2014; Reaburn, 2014; Wasserstein & Lazar, 2016). Some mathematics teachers' interpretation of $p$-values did not demonstrate any conceptual understanding of $p$-values. These teachers' weak understanding of $p$-values led to their pedagogical decision of avoiding conceptual understanding of $p$-values. For instance, RY was very honest about her lack of understanding of the $p$-value. So she did not even try to address its definition. Although many mathematics teachers provided the textbook definition of

$p$-values to students, QK, YW, and FX stated that students found the formal definition of $p$-values difficult to comprehend. As a result, these teachers' instruction of $p$-values merely focused on applying $p$-values to make decisions in hypothesis testing. Their students were not required to understand what a $p$-value means or what it measures. They only needed to memorize rules for exams.

Among those who demonstrated some conceptual understanding of $p$-values, many teachers attempted to address the $p$-value beyond its textbook definition and its application in hypothesis testing, writing an alternative definition of $p$-values and making a connection to other concepts in statistics. However, because their understanding of $p$-values was incomplete, many of those alternative definitions of $p$-values were incorrect or misleading. The noticeable exception was DE, who revealed his solid understanding of the $p$-value by not only providing an appropriate alternative definition, but also utilizing modern technology and simulation to present $p$-values visually to his students prior to the formal introduction of the concept. Perhaps formal statistics education is imperative for mathematics teachers to teach the concept of $p$-values.

**Delivery Method**

In mathematics, there is usually more than one approach to the final answer. In statistics, there is also more than one way to present an idea. Participating mathematics teachers with different thinkings tended to deliver the same concept differently. For example, in item 6, mathematics teachers presented five different ways to address the effects of outliers. On the one hand, teachers with strong mathematical thinking chose to explain the effects numerically by computing descriptive statistics from a dataset before and after the inclusion of

outliers. Some teachers brought datasets to life by showing their real-world applications, a method that also demonstrated statistical thinking in context. On the other hand, teachers with strong statistical thinking used visual representation, with or without technology, to illustrate numerical change and visual change in descriptive statistics before and after adding an outlier.

When addressing the consistency in categorical data, many mathematics teachers revealed their misconceptions about comparing consistency using frequency. For example, DE thought a consistency in categorical data means a minimal difference in bar heights of a bar graph. This misconception is commonly caused by viewing the bar graph without considering the meaning of the bars. A bar graph's vertical axis represents the frequency of each category, not the actual data value. DE and many other mathematics teachers who chose grade 7 were addressing the consistency in frequency, not in data. Among a few other mathematics teachers who answered correctly, KU and HH recognized those bars' meaning in the bar graph. They were able to justify their answers by transforming the bar graph to another form, such as a frequency table or the original list of data. In addressing a subsequent pedagogical question, RM also realized that, by implementing technology and transforming the bar graph to a different representation, it would be much easier for students to understand why grade 8, not grade 7, had the most consistent response.

Furthermore, some mathematics teachers went beyond the textbook when presenting a statistical concept. In item 3, YW's response demonstrated her mathematical thinking on data collection and her additive view of a sample. In her teaching on random sampling, YW stuck to the textbook definition of a random sample and stressed that the theoretical probability of every member being chosen from the population had to be the same in order for the sample to be random. Unlike other mathematics teachers who used real-life examples to illustrate what a

random sample means, YW did not provide any other examples. According to studies (Saldanha & Thompson, 2002; Watson, 2013), students who were taught only the theoretical aspect of random sampling might view a sample independent from the population it was drawn from, which could hinder their understanding of sampling distribution and inferential statistics. To avoid this potential misconception of students', RY, who demonstrated statistical thinking on data collection and a multiplicative view of a sample, applied real-life examples to illustrate the uncertainty in a random sampling process and deemphasized the theoretical equality among all members of the population. By her logic, a random sample should be constructed in such a way that not only members of the sample should all come from the population, but also the characteristics of the sample would, to some degree, mirror the same characteristics of the original population. What's more, RY recommended using technology for a virtual simulation of activities such as coin tossing or card drawing. Students can explore the concept of random sampling and learn from hands-on experience.

**Student Assessment**

Those teachers with primarily mathematical thinking were more likely to assess students' mathematical knowledge in the form of computation. Correspondingly, the teachers with primarily statistical thinking preferred evaluating students' statistical knowledge, which involved various representations of statistical concepts and contextual interpretation of statistical outputs. For instance, many mathematics teachers were able to provide the mathematical meaning of $p$-values in item 9 and interpret the regression equation's algebraic meaning in item 10. QK, YW, and FX stated that their assessments on $p$-values and linear regression focused on the rejection rules concerning $p$-values and the procedure used to find

the regression equation by hand. Contrarily, when evaluating students' knowledge in inferential statistics, DE, RM, and HH, who interpreted the statistical meaning of $p$-value via online simulations, paid more attention to students' visual understanding of $p$-values in different types of hypothesis tests. In particular, RM required her students to visualize the $p$-value for every problem in tests of significance. In terms of linear regression, all three of them considered it a necessity for students to know how to use technology to conduct linear regression and interpret results. Notably, HH graded students' work in linear regression based on their ability to interpret instead of their ability to calculate. She also used technology to test students' knowledge of new topics, such as residual plot and model fitting, that were not mentioned by other mathematics teachers.

# Chapter 7

# RESULTS: RESEARCH QUESTION 3

According to Groth (2013), by switching their perspective on statistics from themselves to students, it is possible for statistics instructors to transform their own key developmental understandings (KDUs) in statistics into "pedagogically powerful ideas" or pedagogical content knowledge that significantly advances students' conceptual understanding in statistics. This chapter identified these powerful ideas in the teachers' responses during the interview and provided responses to the third research question: *With a general knowledge of various types of statistical technology options, how do mathematics teachers promote statistical learning through teaching with or without technology?* In each subsequent section, examples were given to reflect the two subdomains in pedagogical content knowledge: awareness of common student difficulties in statistics and selecting teaching strategies unique to statistics.

## Statistics and Students

There were powerful ideas from the 15 mathematics teachers that were pedagogical decisions made based on students' learning experiences in statistics. Some of them were inspired by mathematics teachers' own learning experiences when they were students. But many of them came from years and years of classroom teaching experience.

To start with, mathematics teachers might lower the difficulty of a task in statistics when most students could not complete the task with what they had learned. Semester projects are popular among many introductory statistics courses. Some projects can be very open-ended and involve all four statistical problem-solving components: formulating

questions, collecting data, analyzing data, and interpreting results. When QK first started teaching statistics, he implemented a project that asked students to conduct a statistical analysis from scratch. But soon, he realized that neither his students nor he had the knowledge to complete the project. In his subsequent semesters of teaching, QK learned that it was not easy for introductory statistics students to develop good, sound statistical questions that provide meaningful learning experiences. When students worked on formulating questions, they had only begun their statistical journey–they had not learned about the other three components of a typical statistical analysis yet. However, decision-making in one component depends on the decision-making in the other three components; they influence each other, and they rely on each other. With such a realization, QK provided a list of pre-approved statistical questions for students to choose from, making the project something that his students could accomplish using their existing knowledge. Moreover, because all questions were pre-made, QK was able to anticipate difficulties that students might encounter in subsequent components and be better prepared to assist his students. Similarly, HH commented that a project based on option C in item 2 might look easy but could be very hard for students to implement. Students typically had a hard time obtaining a decent sample that avoids biases. Thus HH directed her students to find an existing large dataset from the Census at School. To test students' knowledge of sampling methods, she asked students to apply a random sampling method to draw a smaller sample from the dataset they found.

In addition, mathematics teachers might place extra emphasis on a certain topic to prepare their students for a later topic in statistics. Based on the responses from item 9, quite a few mathematics teachers reported that the $p$-value was a commonly known thorny concept to many statistics students, regardless of their type of thinking in statistics. As a matter of fact,

even some participating teachers demonstrated common misconceptions when interpreting $p$-values. Some mathematics teachers with solid statistical understanding about $p$-values, such as WG and HH, claimed that understanding probability was the key to interpreting $p$-values correctly. According to them, probability was closely related to the concept of $p$-values. Helping students understand probability could resolve many students' difficulties in learning $p$-values. For instance, WG stated that sampling distribution depended on the law of large numbers, and the decision-making in hypothesis testing essentially borrowed the idea from the rare event rule. More importantly, both sampling distribution and hypothesis testing were essential to understanding $p$-values visually and contextually.

Lastly, mathematics teachers could occasionally assign procedural tasks that did not involve statistical thinking. This may sound like the opposite of a "powerful idea" at first, but it is a pedagogical choice made by experienced teachers who care about students' learning and growth. For example, when addressing their opinion about hand calculating $r$, RY, KU, and AQ stated that the successful experience of computing something as complicated as $r$ could create a sense of achievement for students with non-STEM backgrounds. These tasks aimed to boost the confidence of students with low self-esteem and improve their learning experiences in mathematics and statistics. Ideally, after completing these tasks, students would believe in themselves again and be prepared to learn something more challenging and conceptual in statistics.

### Statistics and Teaching

A good understanding of statistics involves a grasp of many intricate concepts and an ability to apply statistical thinking, which differs from mathematical thinking. For mathematics

teachers, this presents a challenge. In order to be able to teach statistics and statistical thinking, the teacher needs to know not only the statistical interpretation of the concept but also how certain statistical concepts should be taught so that students can perceive statistics differently from mathematics.

Firstly, many concepts in statistics can be explained more efficiently by attaching real-life contexts relevant to students. According to RM, she usually grabbed students' attention by analyzing their real grades. In item 3, RM proposed three plans to give out extra credits:

1. Give extra credit to students whose last name starts with M
2. Give extra credit to students who scored 100 on exams
3. Give extra credit to students randomly selected by a random number generator

Initially, most students claimed that plan 2 and plan 3 are fair. After RM changed plan 2 to "give extra credit to students who scored 60 and below on exams," many students changed their minds about plan 2 and stated that it was no longer fair. In the end, RM explained that neither the original plan 2 nor the new plan 2 was fair because plan 2 did not give everyone an equal chance to receive extra credit. Through this exercise, RM explained the meaning of fairness in statistics and pointed out how it might be used differently in other contexts.

In item 6, RM gave us another great example of attaching real-life contexts–using students' Blackboard grades to illustrate the effects of outliers or extreme values on the mean and median. There were a few students in RM's class who missed an exam and scored zero. Blackboard automatically calculated both the mean exam score and the median exam score of all students. RM told her students that Blackboard's mean exam score was biased and underestimated the central tendency of all exam scores. The median exam score was a better measure of a typical exam score because it was less affected by zero scores. As a result, RM

informed her students that they should use the median exam score to gauge their class performance.

Secondly, mathematics teachers should realize that data in statistics can be presented in various forms. Some forms facilitate understanding, but some forms deceive viewers. Recall that item 5 asked participants to select the grade level with the highest consistency based on a bar graph. Unfortunately, many mathematics teachers incorrectly addressed the bar graph's consistency and selected the grade that had the lowest consistency. They mixed up the actual data with the frequency counts. To clarify the solution, RM and HH applied transnumeration in statistics and transformed the bar graph into a table form and a list form. The table form was simply a frequency table with responses in the first column and their frequency counts in the second column. The list form was the original dataset, which contained only the two response choices–the zoo and the aquarium. The table form and the list form made it clear that grade 7 had evenly distributed responses between the zoo and the aquarium, so it was the least consistent. On the other hand, grade 8 had the highest frequency count for the aquarium, indicating that most eighth-graders agreed to go to the aquarium.

However, one should always be extra cautious when applying alternative representations in statistics. Sometimes, they may lead to misconceptions. For example, in item 8, KU used an intuitive method to illustrate the effect of the confidence level on the width of confidence intervals. In her illustration, KU asked her students to give an interval of any width to estimate her age. Her students responded with intervals of various widths, some very wide and some very narrow. KU selected three such intervals, one very wide, one very narrow, and one of normal width, asking her students which interval they felt most confident with. Naturally, her students picked the wide interval for its inclusion of more possibilities. Then KU

stated that as the confidence level increased, the width of the confidence interval became wider. This illustration may sound like a powerful idea that enables students to easily remember the effect of the confidence level on the width of confidence intervals. But, KU's method incorrectly interpreted the confidence level as one's confidence in the confidence interval being accurate. To make this idea close to the true interpretation of the confidence interval, KU could rephrase her question in this way: which interval do you think will yield a procedure that has the highest chance of capturing the true age of mine? However, this question requires students' knowledge of probability and a solid understanding of the confidence level which makes the illustration less intuitive.

Lastly, technology should be used to demonstrate complex statistical concepts, especially those that rely on the notion of repeated sampling. During the interview, DE mentioned on multiple occasions his use of StatKey and online simulation to help him clarify statistical topics. DE discussed his use of computer simulation to demonstrate sampling variability in item 3, the effects of outliers in item 6, the statistical interpretation of the confidence level in item 8, and visualization of $p$-values in item 9. Based on DE's description, his students would normally explore a given concept via technology first and then receive the proper formal introduction to the concept. DE claimed that most textbook examples were created based on a single sample. He would like his students to contemplate what happens if the same experiment was repeated on different samples.

According to studies (Bakker et al., 2006; Cobb & Moore, 1997; Gil & Ben-Zvi, 2011; Konold & Harradine, 2014; Makar, 2014), it is important for students to create their own informal definitions through exploratory and informal analysis in statistics. The exploration could be as simple as an opening question on why the statistical concept matters in real-life or

as complicated as a computer simulation on a sampling distribution. It is both possible and acceptable for the informal definition constructed by students to be incorrect or incomplete. Later, when students are given the formal definition of the concept, they will probably notice something different from their own understanding. At that moment, students will form two definitions of the same concept in their mind–one informal, created by them, and one formal, given by the instructor. But neither definition will be accepted by students. Their informal definition is contrasted by the formal definition, and the formal definition is still new to them. While they receive more instruction on the formal definition and practice with more examples that use it, students will gradually gain new understandings about both definitions in their minds. As these new understandings accumulate, students will eventually attempt to modify their own informal definition accounting for any difference that they have detected. Through such a process, students have not only reinforced their conceptual understanding of statistics but also obtained an interpretation of the concept using their own words and images, as well as connections to other knowledge that they have previously acquired.

# Chapter 8

# SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

## Summary

To elicit and decipher college mathematics teachers' knowledge of statistical thinking, statistical teaching, and statistical technology, 15 part-time statistics instructors (9 females and 6 males) from a local community college in the New York City area participated in a qualitative research study. The study investigated differences between mathematical thinking and statistical thinking when teaching college-level introductory statistics as well as the effects on teaching statistics when mathematics teachers apply different types of thinking. The three research questions dealt with types of thinking when mathematics teachers solve statistical problems, pedagogical choices that are influenced by different types of thinking, and pedagogical ideas that advance students' conceptual understanding of statistics.

Participants went through two online surveys and one in-person interview. In the first online survey, participants reported their academic background and teaching experience. Additionally, as part of the survey questions, participants freely explored and informally evaluated five different technology options for teaching statistics. These technology options included Minitab, Tinkerplots, an Online Statistics Course in Google Sheets, Rossman/Chance Applet Collection, and Rice Virtual Lab in Statistics. The options were selected based on the classifications by Chance et al. (2007):

1. Statistical Software Package: computer programs created for professional statistical analysis.

2. Educational Software: computer programs created for teaching statistics.
3. Spreadsheets: Excel or Google Sheet style programs widely used in finance and business.
4. Web Applets: web applications that typically run in a web browser and allow users to manipulate data in a visual, dynamic, and exploratory manner.
5. Multimedia Materials: computer programs or websites that combine text, audio, images, animation, video, quizzes, and other interactive media in introductory statistics.

The second online survey collected participants' self-reported assessment in technological pedagogical content knowledge (TPACK) through 28 statements across the seven domains of the TPACK framework: Technological Knowledge, Content Knowledge, Pedagogical Knowledge, Pedagogical Content Knowledge, Technological Content Knowledge, Technological Pedagogical Knowledge, and Technological Pedagogical Content Knowledge. For each statement, participants scored 5 points for choosing "strongly agree," 4 points for "agree," 3 points for "neither agree nor disagree," 2 points for "disagree," and 1 point for "strongly disagree." The instrument was adapted from the survey created by the inventor of the TPACK framework (Schmidt et al., 2009).

During the in-person interviews, participants answered nine statistical problems related to four investigative steps in statistical analysis: formulating questions, collecting data, analyzing data, and interpreting results. These statistical problems were selected from the LOCUS assessment, whose main objective is to test students' statistical thinking based on the guidelines for assessment and instruction in statistics education (GAISE) (Franklin et al., 2007). Furthermore, participants addressed pedagogical questions related to the statistical concept in each problem. These pedagogical questions were developed bearing in mind differences between statistical thinking and mathematical thinking in the following aspects:

problem-solving process, variability, context, data production, transnumeration, and probabilistic thinking. Due to COVID-19 and mandatory social distancing when collecting data, only the first five interviews were conducted in the adjunct faculty office where participants worked. The remaining 11 interviews were conducted on video conference software Zoom.

## Conclusions

*Research Question 1: In which ways do statistical thinking and mathematical thinking take place among mathematics teachers when teaching introductory statistics?*

Based on mathematics teachers' responses to nine statistical problems, a framework that hypothesizes two types of mathematics teachers' thinking in statistics was created (Table 5.6). At the beginning stage of a statistical study, developing a question that does not require data collection and expects a deterministic answer is considered mathematical thinking. Formulating a question that requires data collection and evokes variability in response is deemed to be statistical thinking. These two types of questions are defined as the survey question and the statistical question, respectively. For example, according to 15 mathematics teachers' responses to item 2 from the interview, "How many students attend the school?" is a survey question that many mathematics teachers agreed that does not need data collection to address, but "How many text messages do students at the school send per week?" is a statistical question that can be addressed by a random sample of students from that school. Additionally, identifying the correct population of interest from a statistical question is essential for determining samples and considered statistical thinking. The question "Do students at this school have higher test scores than students in other schools in the district?" is a statistical question that compares two different populations and can be addressed by a random sample of

students at the school and a random sample of students from other schools in the district. Similarly, "How many hours does each class at the school meet per year?" studies a characteristic of the school and can be addressed by a random sample of all the classes at the school.

During data collection, collecting a sample that is only part of the population reveals the additive view of a sample and mathematical thinking. To be considered statistical, the sample must be collected with a multiplicative view in which the sample not only comes from the population but also represents the primary characteristics of the population. In item 3 from the interview, a student wants to estimate the mean number of books that have been read by all students at his school over the summer. All mathematics teachers, except one, considered it biased if the student surveys the first 35 students who enter the library or a random sample of students entering the library. Both samples lack the inclusion of students who do not enter the library, which could have omitted crucial data that may lead to very different estimates for the study. In particular, one mathematics teacher revealed her additive view of a sample by selecting a random sample from only those who enter the library–a subset of all students at the school. And one other mathematics teacher revealed her multiplicative view of a sample by stating explicitly that an unbiased sample needs to be drawn from students who enter the library and students who do not enter the library–a mini version of the original population.

When it comes to analyzing data, regardless of the type of thinking, a mathematical model is typically chosen as the theoretical foundation. However, how this model is implemented depends on the type of thinking. Participants were asked two questions in item 7 during the interviews:

Question 1: Assume a coin is fair. If we toss the coin five times, how many heads will we get?
Question 2: You pick up a coin. Is this a fair coin?

Both questions are indeterministic in their final answer, and both questions can be answered by implementing the binomial distribution for the number of heads when tossing a fair coin five times. The first question is mathematical because it can be answered by applying the binomial distribution and theoretical probability knowledge directly. The number of heads could be 0, 1, 2, 3, 4, or 5, with 2 or 3 heads the most likely. Some mathematics teachers responded 2 or 3 heads, a rounded result based on the theoretical answer 2.5 heads, revealing their mathematical thinking when applying the binomial distribution. The second question is statistical since it requires conducting experiments with the coin and calculating empirical probabilities. Many mathematics teachers proposed an experiment in which the inspected coin will be tossed five times in a row. As these tosses are repeated a large number of times, one can construct an empirical distribution for the number of heads when tossing the coin five times. If this distribution is significantly different from the binomial distribution from question 1, then one can conclude that the coin is probably not fair. These mathematics teachers demonstrated their statistical thinking for analyzing a coin's fairness based on generated data. In short, question 1 is solved based on the implementation of the theoretical model only, but question 2 requires both the theoretical model and empirical data that account for data production and context.

In addition to the selection of a mathematical model, data analysis also involves reasoning. Generally speaking, mathematical reasoning adopts mathematical formulas and properties, and statistical reasoning accounts for statistical context and variability. For example, during the interview, mathematics teachers rationalized why, as sample size

increases, the width of confidence interval increases in item 8. Many mathematics teachers applied the mathematical formula of the margin of error. They used the mathematical fact that, in a fraction, when the numerator is fixed, the fraction decreases as the denominator increases. One mathematics teacher stated that, as sample size increases, sampling error would decline, lowering the maximum likely size of sampling error, or the margin of error, resulting in a narrower confidence interval. This reasoning connected the margin of error to the sampling error and interpreted the concept of margin of error from a statistical point of view.

Lastly, if results are reported from the sample with no inference about the population, it applies mathematical thinking. If a conjecture about the population is identified but written in words that consist of formal definition or rules from the textbook, it still applies primarily mathematical thinking. If results are interpreted in terms that infer contextual generalization about the population and indicate variability in data, it applies statistical thinking. For example, the following interpretations of the 95% confidence interval from statistical problem item 8 correspond to the aforementioned three cases, respectively.

1. The 95% confidence interval is $[-0.11, 0.27]$.
2. With 95% confidence, the difference in the proportion of females and males at the school who have taken an honors class could be between $-0.11$ and 0.27.
3. Based on the sample, the difference in the proportion of females and males at the school who have taken an honors class could be between $-0.11$ and 0.27. If this procedure is to be repeated on all possible samples from the population, the proportion of all computed confidence intervals that will capture the true proportion difference would be approximately 95%. Since this CI includes 0, there may not be a difference between the proportions of the two genders.

Even though context plays an important role in statistical thinking, it does not automatically indicate statistical thinking. For instance, when participants addressed the number of heads when tossing a fair coin five times, five participants responded 2 or 3 heads. They realized that the context for the number of heads required the answer to be an integer. Since the expected number of heads out of five tosses is 2.5, they rounded the answer to 2 and 3. This numerical answer amendment is essentially a reflection of their mathematical thinking. These five participants did not accept the theoretically accurate 2.5 heads because it was practically impossible to obtain 2.5 heads when tossing a fair coin five times in reality, not because they applied any statistical thinking. It commonly occurs in algebra when an equation created from a word problem yields a positive solution and a negative solution, but only the positive solution was kept in the end because the sought quantity could not be negative by context.

Similarly, though many responses involving mathematical procedures or numerical properties in this study were categorized as mathematical thinking, some of them also revealed participants' statistical thinking. For example, in item 2, three mathematics teachers recommended the mathematical recipe of applying the confidence interval to estimate the population's mean number of texts sent per week. They applied a standard statistical method, the confidence interval for the mean, based on a mathematical model. But they also initiated the model selection, which required their statistical understanding of the problem in order to determine the most appropriate statistical method to implement. Many practice problems in a typical introductory statistics textbook comprise descriptive statistics and a specific statistical method that students should implement to solve the problem. They feed the rote learning of

statistical procedures and hinder students' independent thinking when conducting a statistical analysis.

It is worth mentioning that the aforementioned differences between statistical thinking and mathematical thinking were categorized based on the specific task described and the primary knowledge required. In some aspects, the boundary that differentiates the two types of thinking is not clearly defined. For example, applying a statistical method to solve a problem could be interpreted as mathematical thinking for its application of a mathematical model and mathematical formulas. But it could also be interpreted as statistical thinking for the judgment that one has to make to select an appropriate statistical method to solve the problem. In other words, the decision-making in differentiating statistical thinking from mathematical thinking tends to be nonbinary in some cases.

*Research Question 2: With a general knowledge of various types of statistical technology options, how does mathematics teachers' statistical thinking or mathematical thinking affect their way of teaching?*

The first area in which mathematics teachers' thinking affects their teaching is topic coverage in statistics. Participating mathematics teachers covered topics that they believed they understood. For topics that they were not familiar with, depending on whether or not the topic would be tested on the exam, many mathematics teachers either chose to skip or cover at a level that they felt they could handle. For example, many mathematics teachers lacked a complete understanding of the concepts in the statistical question and the survey question during the interview. They did not teach the concept of the statistical question. Among those four who did teach it, only one mathematics teacher assessed students' conceptual knowledge of the statistical question by integrating the concept into a semester-long project.

Those teachers with strong mathematical thinking tended to cover more mathematical topics. Correspondingly, those teachers with strong statistical thinking were inclined to cover more statistical issues. For instance, mathematics teachers who mostly applied mathematical knowledge to address statistical problems in the interview agreed that hand computation of the correlation coefficient $r$ is a necessary experience for students studying statistics. It enables students to unfold mathematical steps that lead to statistical results and appreciate statistical outputs produced by technology more. On the contrary, some mathematics teachers who generally applied statistical knowledge in the interview believed that it is more important for students to develop the ability to interpret descriptive statistics and inferential statistics. Numerical calculations should be done using modern technology. It prepares students to be more informed citizens as well as better job applicants in contemporary society.

Probability is a topic both mathematical and statistical. As a result, participating mathematics teachers held very different opinions about what topics in probability should be instructed in an introductory statistics course, especially among those who demonstrated mathematical thinking in the interview. However, those mathematics teachers who demonstrated statistical thinking in statistical problem item 7 reached a consensus that, as a minimum, fundamental topics in probability should always be taught. Many statistical concepts in inferential statistics depend on the understanding of basic probability.

When it comes to the delivery method, the teachers with different thinkings tended to present the same statistical concept differently. For instance, when describing the effects of outliers on descriptive statistics, the participants with mathematical thinking presented the numerical change in descriptive statistics before and after the inclusion of outliers. In contrast, the teachers with statistical thinking focused on the distribution of the dataset. They illustrated

both the numerical change and visual change in descriptive statistics before and after adding an outlier. While teaching multiple representations of the same concept typically facilitates students' conceptual understanding, extra caution should be placed to avoid misconceptions caused by incorrect alternative representations.

Lastly, the teachers with primarily mathematical thinking leaned towards assessing students' mathematical knowledge that involved little statistical thinking, such as computation of descriptive statistics, procedural work in hypothesis testing, and applying formulas in simple linear regression. Notably, according to some participants, their students were instructed to memorize rejection rules about $p$-values with little or no conceptual understanding. Those teachers with primarily statistical thinking were more than likely to evaluate students' statistical knowledge by looking for various representations of statistical concepts and contextual interpretations of statistical outputs. For example, some teachers reported that, in addition to numerically computing descriptive statistics, their students were tested on their ability to interpret them in the context of the problem; in addition to following a recipe to conduct a hypothesis test, their students were required to visualize $p$-values and validate their answer; in addition to finding the linear regression equation by hand, their students were instructed to interpret the statistical meaning of the slope and apply it to find the expected change in dependent variables when a difference in independent variables is given.

*Research Question 3: With a general knowledge of various types of statistical technology options, how do mathematics teachers promote statistical learning through teaching with or without technology?*

By switching their perspective on statistics from themselves to students, mathematics teachers become aware of common student difficulties in statistical learning. First,

mathematics teachers circumvent a difficult task in statistics when most students cannot complete the job with what they have learned. They test the original complex topic in an alternative form, so the task's objective remains the same. For example, instead of asking his students to create a sound statistical question on their own, one mathematics teacher provided students with a list of questions, including both statistical questions and survey questions. Students were instructed to choose the appropriate statistical question they thought could be addressed by collecting a random sample. Second, mathematics teachers place extra emphasis on a certain topic to prepare students for a later topic in statistics. Many participating mathematics teachers were aware of the difficulties statistics students had when learning about $p$-values. To rectify the situation, some mathematics teachers with strong statistical thinking stated that probability was the key. Helping students understand topics in probability, such as the law of large numbers and the rare event rule, could improve many students' understanding of $p$-values. Third, mathematics teachers occasionally assign procedural tasks that involved little statistical thinking. According to participants, most of their students came from non-STEM backgrounds who had terrible experiences learning math and statistics. Guiding these students through procedural tasks that are neither too easy nor too difficult can boost low self-esteemed students' confidence and prepare them for more challenging and conceptual statistical topics.

By switching their perspective on statistics from themselves to students, mathematics teachers present statistics in a way that attempts to distinguish statistics from mathematics and fosters the development of statistical thinking in introductory statistics students. Firstly, many concepts in statistics can be explained more efficiently if they are attached to a real-life context relevant to students. For example, according to one mathematics teacher, many data analyses

in statistics can be performed on students' actual grades so students may perceive statistics as a tool to improve decision making in life, rather than another course that they must go through to fulfill their academic requirements. Secondly, mathematics teachers should realize that data in statistics can be presented in various forms. Sometimes, they facilitate understanding, but some other times they deceive viewers. For instance, using a bar graph is typical for visualizing qualitative data. However, when addressing consistency in categorical data, the bar graph from item 5 during the interview tricked many mathematics teachers into thinking that consistency in data meant consistency in the bars' heights. Two mathematics teachers with strong statistical thinking recommended that transforming the bar graph into a frequency table or its original list of categorical responses could help other teachers clarify the mistake they made. Lastly, technology should be implemented to demonstrate abstract statistical concepts, especially those that rely on the notion of repeated sampling. According to one mathematics teacher who frequently used computer simulations to illustrate statistical concepts, technology promotes exploratory and informal analysis in statistics, making formal instruction more effective and thus advancing students' conceptual understanding of statistics.

## Recommendations

Several improvements can be made to this study for those who are interested in undertaking similar work. First, the sample could be more representative. The sampling method adopted in this study was not random. Only mathematics teachers who were interested in the study or who had time for the study participated in the study. The final sample made evident that all 15 mathematics teachers were part-time instructors at the college. Responses lacked opinions from full-time instructors who might have more insights into

curriculum design choices for introductory statistics as well as standardized statistics exams made by the mathematics department. Because the study was only able to afford the recruitment of 15 mathematics teachers from one college, findings and conclusions drawn from this small sample were limited to a specific group and subject to potential biases. Second, mathematics teachers' pedagogical content knowledge assessment could be more formal. In this study, participants' statistical knowledge was assessed through in-person interviews in which mathematics teachers verbally provided their solution and rationale. Even though the interview handout had left enough white space after each problem, very few mathematics teachers provided any written work. Many mathematics teachers chose to devise a plan for the problems that required a numerical answer without carrying it out. As a result, some analyses were performed based on incomplete statistical work from teachers, which might be misinterpreted. Instead of an interview, mathematics teacher's statistical knowledge could be assessed via formal written work, and their pedagogical knowledge could be evaluated via classroom observations. Third, the technology evaluation survey could be administered more uniformly. During the first online survey, 15 mathematics teachers freely explored and informally ranked the five different technology options. The study did not track how many links each participant clicked or how much time each participant spent exploring each link. Therefore, it is reasonable to speculate that the amount of information gathered about each technology option varied by participants, and their rankings of technology options were probably not comparable. If time is not an issue, the technology evaluation survey should be conducted via in-person interviews where all participants explore the same aspects of each technology option under systematic guidance from the researcher.

People who work in education or a related field may benefit from some results of this study. First, hypothesized aspects of mathematics teachers' statistical thinking and mathematical thinking can be used by the mathematics department's statistics course coordinator as a theoretical framework for creating standardized exams that test students' statistical thinking. Meanwhile, the framework can also be used to design rubrics of a statistics course project that involves four investigative steps of statistical problem-solving. Second, teachers' pedagogical ideas that advance introductory statistics students' conceptual understanding can be borrowed and modified by current statistics instructors to improve the quality of teaching in statistics. Third, some discussions about statistical misconceptions, such as confidence intervals and $p$-values, can be continued and extended by teacher educators who are responsible for cultivating the next generation of statistics instructors. These discussions also highlight the importance of formal statistics education for prospective teachers teaching introductory statistics.

A future study can concentrate on university-level mathematics teachers who are teaching introductory statistics. While most participants in this study were selected from a community college, a few also taught at a local university. Based on their responses, topic coverage in statistics differs between community colleges and universities. At the university level, more topics in inferential statistics are covered. A study focusing on university-level teachers could extend the results from this study and complement the short discussion concerning later topics of introductory statistics. Additionally, a similar study can be conducted focusing on statistics instructors who are not mathematics teachers. It will be interesting to see how instructors coming from a non-mathematical discipline address statistical concepts, statistically, mathematically, and pedagogically. Alternatively, a study can

be designed to investigate different types of thinking between mathematics and other disciplines that also use mathematical knowledge extensively. Furthermore, since mathematics teachers held different opinions about what topics in probability should be covered in an introductory statistics course, a study could be designed and carried out to find topics in probability that are essential for students' conceptual understanding of introductory statistics. It could be a study that assesses students' knowledge in statistics related to probability, both quantitatively and qualitatively, or a study that seeks opinions from experienced statistics instructors and professional statisticians, or a combination of both.

References

Allmond, S., & Makar, K. (2010). Developing primary students' ability to pose questions in statistical investigations. In *Proceedings of the 8th international conference on teaching statistics. voorburg, the netherlands: International statistical institute.*

Arnold, P. (2007). What about the p in the ppdac cycle? an initial look at posing questions for statistical investigation. *Education*, *55*.

Bakker, A. (2002). Route-type and landscape-type software for learning statistical data analysis. In *B. phillips (chief ed.), developing a statistically literate society: Proceedings of the sixth international conference on teaching statistics, voorburg, the netherlands (cd-rom).*

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, *3*(2), 64–83.

Bakker, A., Biehler, R., & Konold, C. (2004). Should young students learn about box plots. *Curricular development in statistics education: International Association for Statistical Education*, 163–173.

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical thinking and learning*, *13*(1-2), 5–26.

Bakker, A., Derry, J., & Konold, C. (2006). Using technology to support diagrammatic reasoning about center and variation. *Proceedings of ICOTS-7. Salvador de Bahia, Brasil: International Association for Statistical Education.*

Bargagliotti, A., Anderson, C., Casey, S., Everson, M., Franklin, C., Gould, R., … Watkins, A. (2014). Project-set materials for the teaching and learning of sampling variability and regression. In *Sustainability in statistics education. proceedings of the ninth international conference on teaching statistics (icots9), flagstaff, arizona, usa. voorburg, the netherlands: International statistical institute.*

Batanero, C., Arteaga, P., & Ruiz, B. (2010). Statistical graphs produced by prospective teachers in comparing two distributions. In *Proceedings of the sixth congress of the european society for research in mathematics education* (pp. 368–377).

Bayarri, M. J., & Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 58–80.

Ben-Zvi, D. (2007). Using wiki to promote collaborative learning in statistics education. *Technology Innovations in Statistics Education*, *1*(1).

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM*, *44*(7), 913–925.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review*, *65*(2), 167–189.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2012). Technology for enhancing statistical reasoning at the school level. In *Third international handbook of mathematics education* (pp. 643–689). Springer.

Birel, G. K., & Çakıroğlu, E. (2018). Preservice mathematics teachers' tpack development in statistics teaching: A microteaching lesson study. In *Contribution paper of tenth international conference on teaching statistics (icots10)*.

Bruno, A., & Espinel, M. (2009). Construction and evaluation of histograms in teacher training. *International journal of mathematical Education in Science and Technology*, *40*(4), 473–493.

Bureau of Labor Statistics. (2019). U.s. department of labor, occupational outlook handbook, mathematicians and statisticians. https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm#tab-6. (Accessed on 08/10/2019).

Burgess, T. (2002). Investigating the "data sense" of preservice teachers. In *Proceedings of the sixth international conference on teaching statistics. cape town, south africa: International association for statistics education. online: Ww. stat. auckland. ac. nz/ĩase/publicatons*.

Cai, J. (2000). Understanding and representing the arithmetic averaging algorithm: An analysis and comparison of us and chinese students' responses. *International Journal of Mathematical Education in Science and Technology*, *31*(6), 839–855.

Cai, J., & Gorowara, C. C. (2002). Teachers' conceptions and constructions of pedagogical representations in teaching arithmetic average. In *Proceedings of the sixth international conference on teaching of statistics, cape town. voorburg, the netherlands: International statistical institute*.

Capraro, M. M., Kulm, G., & Capraro, R. M. (2005). Middle grades: Misconceptions in statistical thinking. *School Science and Mathematics*, *105*(4), 165–174.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., … Witmer, J., et al. (2016). Guidelines for assessment and instruction in statistics education (gaise) college report 2016.

Casey, S. A., & Wasserman, N. H. (2015). Teachers'knowledge about informal line of best fit. *Statistics Education Research Journal*, *14*(1).

Chai, C. S., Koh, J. H. L., & Tsai, C.-C. (2016). 6a review of the quantitative measures of technological pedagogical content knowledge (tpack). In *Handbook of technological pedagogical content knowledge (tpack) for educators* (pp. 97–116). Routledge.

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

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. In *Proceedings of the seventh international conference on teaching statistics* (pp. 1–6). International Statistical Institute Voorburg, The Netherlands.

Chick, H. L., Pfannkuch, M., & Watson, J. M. (2005). Transnumerative thinking: Finding and telling stories within data. *Curriculum matters*, *1*, 87–109.

Chick, H., & Watson, J. (2001). Data representation and interpretation by primary school students working in groups. *Mathematics Education Research Journal*, *13*(2), 91–111.

Chin, C., & Kayalvizhi, G. (2002). Posing problems for open investigations: What questions do pupils ask? *Research in Science & Technological Education*, *20*(2), 269–287.

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

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

Cobb, P. (1994). Where is the mind? constructivist and sociocultural perspectives on mathematical development. *Educational researcher*, *23*(7), 13–20.

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

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Crites, T., & Laurent, R. T. S. (2015). *Putting essential understanding of statistics into practice in grades 9-12*. National Council of Teachers of Mathematics.

del Mas, R. C. (2004). A comparison of mathematical and statistical reasoning. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79–95). Springer.

Dick, T. (2008). Keeping the faith. *Research on technology and the teaching and learning of mathematics: Cases and perspectives*, *2*(2), 333.

Dierdorp, A., Bakker, A., Eijkelhof, H., & van Maanen, J. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning*, *13*(1-2), 132–151.

Ertmer, P. A. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration? *Educational technology research and development*, *53*(4), 25–39.

Espinel, C., Bruno, A., & Plasencia, I. (2008). Statistical graphs in the training of teachers. *C. Batanero; G. Burrill; R. Reading; A. Rossman.(2008). Proceedings of the Joint ICMI/IASE Study Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education.*

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(1), 27–37.

Fielding-Wells, J. (2010). Linking problems, conclusions and evidence: Primary students' early experiences of planning statistical investigations. In *Proceedings of the 8th international conference on teaching statistics. voorburg, the netherlands: International statistical institute.*

Fielding-Wells, J., & Makar, K. (2015). Inferring to a model: Using inquiry-based argumentation to challenge young children's expectations of equally likely outcomes.

Fisher, R. (1946). *Statistical methods for research workers*. Biological monographs and manuals. Oliver and Boyd. Retrieved from https://books.google.com/books?id=Z7keAQAAIAAJ

Fitzallen, N. E. (2013). Characterising students' interaction with tinkerplots. *Technology Innovations in Statistics Education*, *7*(1).

Fitzallen, N., Watson, J., & English, L. (2015). Assessing statistical inquiry. In *The 39th conference of the international group for the psychology of mathematics education* (Vol. 2, pp. 305–312).

Forster, P. A. (2006). Assessing technology-based approaches for teaching and learning mathematics. *International Journal of Mathematical Education in Science and Technology*, *37*(2), 145–164.

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

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (gaise) report. *Alexandria: American Statistical Association.*

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Science & Business Media.

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

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *Bmj*, *327*(7417), 741–744.

Gil, E., & Ben-Zvi, D. (2010). Emergence of reasoning about sampling among young students in the context of informal inferential reasoning. In *Eighth international conference on teaching statistics (icots 8), ljubljana, slovenia*.

Gil, E., & Ben-Zvi, D. (2011). Explanations and context in the emergence of students' informal inferential reasoning. *Mathematical Thinking and Learning*, *13*(1-2), 87–108.

González, M. T., Espinel, M. C., & Ainley, J. (2011). Teachers' graphical competence. In *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 187–197). Springer.

González, M., & Pinto, J. (2008). Conceptions of four pre-service teachers on graphical representation. *Joint ICMI/IASE study: Teaching statistics in school mathematics. Challenges for teaching and teacher education. Proceedings of the ICMI Study*, *18*.

Good, P. I. (2006). *Resampling methods*. Springer.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. In *Seminars in hematology* (Vol. 45, *3*, pp. 135–140). Elsevier.

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

Groth, R. E. (2013). Characterizing key developmental understandings and pedagogically powerful ideas within a statistical knowledge for teaching framework. *Mathematical Thinking and Learning*, *15*(2), 121–145. doi:10.1080/10986065.2013.770718

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning*, *8*(1), 37–63.

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

Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education research journal*, *3*(2), 17–41.

Hand, D. J. (1998). Breaking misconceptions—statistics and its relationship to mathematics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(2), 245–250.

Hannigan, A., Gill, O., & Leavy, A. M. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education*, *16*(6), 427–449.

Henriques, A., & Gutiérrez-Fallas, L. F. (2017). Prospective mathematics teachers' beliefs and tpack for teaching statistics. In *Proceedings of inted2017 conference* (pp. 7193–7203).

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from california's mathematics professional development institutes. *Journal for research in mathematics education*, 330–351.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, *21*(5), 1157–1164.

Hooper, S., & Rieber, L. P. (1995). Teaching with technology. *Teaching: Theory into practice*, *2013*, 154–170.

Jacobbe, T. (2012). Elementary school teachers'understanding of the mean and median. *International Journal of Science and Mathematics Education*, *10*(5), 1143–1161.

Jacobbe, T., & Horton, R. M. (2010). Elementary school teachers'comprehension of data displays. *Statistics Education Research Journal*, *9*(1).

Jacobs, V. R. (1999). How do students think about statistical sampling before instruction? *Mathematics Teaching in the Middle School*, *5*(4), 240.

Kalinowski, P. et al. (2010). Identifying misconceptions about confidence intervals. In *Proceedings of the eighth international conference on teaching statistics* (Vol. 50). Citeseer.

Koehler, M., & Mishra, P. (2009). What is technological pedagogical content knowledge (tpack)? *Contemporary issues in technology and teacher education*, *9*(1), 60–70.

Konold, C. (2007). Designing a data analysis tool for learners. *Thinking with data*, 267–291.

Konold, C., & Harradine, A. (2014). Contexts for highlighting signal and noise. In *Mit werkzeugen mathematik und stochastik lernen–using tools for learning mathematics and statistics* (pp. 237–250). Springer.

Konold, C., & Higgins, T. (2003). Reasoning about data. *A research companion to principles and standards for school mathematics*, *193215*.

Lavigne, N. C., & Lajoie, S. P. (2007). Statistical reasoning of middle school children engaged in survey inquiry. *Contemporary Educational Psychology*, *32*(4), 630–666.

Leavy, A. M., Hannigan, A., & Fitzmaurice, O. (2013). If you're doubting yourself then, what's the fun in that? an exploration of why prospective secondary mathematics teachers perceive statistics as difficult. *Journal of Statistics Education*, *21*(3).

Lee, H. S., & Hollebrands, K. F. (2011). Characterising and developing teachers' knowledge for teaching statistics with technology. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint icmi/iase study: The 18th icmi study* (pp. 359–369). doi:10.1007/978-94-007-1131-0_34

Lehrer, R., Kim, M.-j., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning*, *12*(3), 195–216.

Ling, D. L. (2018). Introduction to statistics using google sheets. http://www.comfsm.fm/~dleeling/statistics/text6.html. (Accessed on 08/11/2019).

Madden, S. R. (2008). Dynamic technology scaffolding: A design principle with potential to support statistical conceptual understanding. In *11th international congress on mathematics education, monterrey, mexico*. Citeseer.

Madden, S. R. (2011). Statistically, technologically, and contextually provocative tasks: Supporting teachers' informal inferential reasoning. *Mathematical Thinking and Learning*, *13*(1-2), 109–131.

Makar, K. (2014). Young children's explorations of average through informal inferential reasoning. *Educational Studies in Mathematics*, *86*(1), 61–78.

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. Taylor & Francis.

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353–373). Springer.

Meletiou-Mavrotheris, M., & Paparistodemou, E. (2015). Developing students' reasoning about samples and sampling in the context of informal inferences. *Educational Studies in Mathematics*, *88*(3), 385–404.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 20–39.

Monk, S. (2003). Representation in school mathematics: Learning to graph and graphing to learn. *A research companion to principles and standards for school mathematics*, 250–262.

Monteiro, C., & Ainley, J. (2006). Student teachers interpreting media graphs. In *Proceedings of the seventh international conference on teaching statistics* (pp. 1–6).

Monteiro, C., & Ainley, J. (2007). Investigating the interpretation of media graphs among student teachers. *International Electronic Journal of Mathematics Education*, *2*(3), 187–207.

Moore, D. S. (1988). Should mathematicians teach statistics?. *College Mathematics Journal*, *19*(1), 3–7.

Moritz, J. (2004). Reasoning about covariation. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227–255). Springer.

Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *Journal of Pharmacology and Experimental Therapeutics*, *351*(1), 200–205.

Oaks, M. (1986). Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

Pea, R. D. (1987). Cognitive technologies for mathematics education. *Cognitive science and mathematics education*, 89–122.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, *12*(3), 30–32.

Pfannkuch, M. (2005). Thinking tools and variation. *Statistics Education Research Journal*, *4*(1), 83–91.

Pfannkuch, M., Arnold, P., & Wild, C. J. (2015). What i see is not quite the way it really is: Students' emergent reasoning about sampling variability. *Educational Studies in Mathematics*, *88*(3), 343–360.

Pfannkuch, M., Budgett, S., Fewster, R., Fitch, M., Pattenwise, S., Wild, C., & Ziedins, I. (2016). Probability modeling and thinking: What can we learn from practice? *Statistics Education Research Journal*, *15*(2).

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17–46). Springer.

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics*, *12*(2), 191–204.

Polya, G. (1971). *How to solve it*. Princeton University Press. Retrieved from http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20%5C&path=ASIN/0691023565

Puiu, T. (2019). Smartphone is millions of times faster than nasa's 1960s computers. https://www.zmescience.com/research/technology/smartphone-power-compared-to-apollo-432/. (Accessed on 08/10/2019).

Reaburn, R. (2014). Introductory statistics course tertiary students'understanding of p-values. *Statistics Education Research Journal*, *13*(1).

Rosenberg, J. (2012). Assessing teachers' tpack. http://matt-koehler.com/tpack2/assessing-teachers-tpack/. (Accessed on 08/11/2019).

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *WIREs Computational Statistics*, *6*(4), 211–221. doi:10.1002/wics.1302

Rouan, O. (2002). Secondary school math teachers' conceptions of the statistical graphics functions, reading and interpretation. In *Proceedings of the sixth international conference on teaching statistics* (pp. 7–12). International Statistical Institute and International Association for ...

Ruggiero, D., & Mong, C. J. (2015). The teacher technology integration experience: Practice and reflection in the classroom. *Journal of Information Technology Education*, *14*.

Saldanha, L., & McAllister, M. (2014). Using re-sampling and sampling variability in an applied context as a basis for making statistical inferences with confidence. In *Sustainability in statistics education. proceedings of the ninth international conference on teaching statistics (icots-9). voorburg, the netherlands: International statistical institute and international association for statistical education.*

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

Santos, R., & da Ponte, J. P. (2014). Learning and teaching statistical investigations: A case study of a prospective teacher. In *Sustainability in statistics education (proceedings of the 9th international conference on the teaching of statistics, flagstaff, arizona, july 13–18). voorburg, the netherlands: International statistical institute.*

Schmidt, D. A., Baran, E., Thompson, A. D., Koehler, M. J., Mishra, P., & Shin, T. (2009). Survey of preservice teachers' knowledge of teaching and technology. (Accessed on 08/09/2019).

Schoenfeld, A. H. (1981). Episodes and executive decisions in mathematical problem solving.

Shaughnessy, J. M., & Pfannkuch, M. (2002). How faithful is old faithful. *Mathematics Teacher*, *95*(4), 252–259.

Shaughnessy, J. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Chap. 21, pp. 957–1009). Charlotte, NC: Information Age.

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

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard educational review*, *57*(1), 1–23.

Sorto, M. A., White, A., & Lesser, L. M. (2011). Understanding student attempts to find a line of fit. *Teaching Statistics*, *33*(2), 49–52.

the Society for Human Resource Management (SHRM). (2016). Jobs of the future: Data analysis skills. https : / / www . shrm . org / hr - today / trends - and - forecasting / research - and - surveys/Documents/Data-Analysis-Skills.pdf. (Accessed on 08/10/2019).

Verillon, P., & Rabardel, P. (1995). Cognition and artifacts: A contribution to the study of though in relation to instrumented activity. *European journal of psychology of education*, 77–101.

Wasserstein, R. L., & Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. Taylor & Francis.

Watson, J. M. (2013). *Statistical literacy at school: Growth and goals*. Routledge.

Watson, J. M. (2017). Linking science and statistics: Curriculum expectations in three countries. *International Journal of Science and Mathematics Education*, *15*(6), 1057–1073.

Watson, J. M., & English, L. D. (2015). Introducing the practice of statistics: Are we environmentally friendly? *Mathematics Education Research Journal*, *27*(4), 585–613.

Watson, J. M., Fitzallen, N. E., Wilson, K. G., & Creed, J. F. (2008). The representational value of hats. *Mathematics Teaching in the Middle School*, *14*(1), 4–10.

Watson, J., Chick, H., & Callingham, R. (2014). Average: The juxtaposition of procedure and context. *Mathematics Education Research Journal*, *26*(3), 477–502.

Watson, J., & Donne, J. (2009). Tinkerplots as a research tool to explore student understanding. *Technology Innovations in Statistics Education*, *3*(1).

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

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). What is statistics? In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 5–36). doi:10.1007/978-3-319-66195-7_1

Willermark, S. (2018). Technological pedagogical and content knowledge: A review of empirical studies published from 2011 to 2016. *Journal of Educational Computing Research*, *56*(3), 315–343.

Yu, C. H. et al. (1995). Identification of misconceptions in the central limit theorem and related concepts and evaluation of computer media as a remedial tool.

Zbiek, R. M., Heid, M. K., Blume, G. W., & Dick, T. P. (2007). Research on technology in mathematics education: A perspective of constructs. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Chap. 27, pp. 1169–1207). Charlotte, NC: Information Age.

Appendix A

Technology Evaluation Survey

**Participant Information**

**Demographics**

1. Your Study ID (provided with your survey link) [SHORT TEXT RESPONSE]
2. Name [SHORT TEXT RESPONSE]
3. Gender [M / F]
4. Age [MULTIPLE CHOICE]

☐ 20-29 ☐ 30-39 ☐ 40-49 ☐ $\geq$ 50

**Academic Background**

1. Highest degree earned: [MULTIPLE CHOICE]

☐ High School ☐ BA ☐ MA ☐ PhD
☐ BS ☐ MS ☐ EdD ☐ n/a

2. Undergraduate Major [SHORT TEXT RESPONSE]
3. Undergraduate Minor [SHORT TEXT RESPONSE]
4. Undergraduate GPA [MULTIPLE CHOICE]

☐ 4.00 ☐ 3.50-3.99 ☐ 3.00-3.49 ☐ $<$ 3.00 ☐ n/a

5. Undergraduate degree received from (country) [SHORT TEXT RESPONSE]
6. Graduate Major [SHORT TEXT RESPONSE]
7. Graduate GPA [MULTIPLE CHOICE]

☐ 4.00 ☐ 3.50-3.99 ☐ 3.00-3.49 ☐ $<$ 3.00 ☐ n/a

8. Graduate degree received from (country) [SHORT TEXT RESPONSE]
9. How many statistics courses have you taken at undergraduate and graduate levels? [SHORT TEXT RESPONSE]
10. How many mathematics courses have you taken at undergraduate and graduate levels? [SHORT TEXT RESPONSE]
11. How many computer science courses have you taken at undergraduate and graduate levels? [SHORT TEXT RESPONSE]
12. How many education courses have you taken at undergraduate and graduate levels? [SHORT TEXT RESPONSE]
13. What other languages do you speak other than English? [SHORT TEXT RESPONSE]

**Teaching Experience**

1. Total years of teaching experience [SHORT TEXT RESPONSE]
2. Age group(s) taught [MULTIPLE CHOICE, MULTIPLE SELECTIONS]
   - ☐ Preschool (< 6)
   - ☐ Elementary (6-11)
   - ☐ Intermedia (11-13)
   - ☐ High school (14-18)
   - ☐ College (19+)
   - ☐ Continuation program for adults (19+)
3. Is your current (main) institution public or private? [PUBLIC / PRIVATE / UNSURE]
4. How many years have you taught statistics at that institution? [SHORT TEXT RESPONSE]
5. What statistics courses have you taught (provide a brief course title)? [SHORT TEXT RESPONSE]
6. What mathematics courses have you taught (provide a brief course title)? [SHORT TEXT RESPONSE]
7. Have you taught mathematics or statistics outside the U.S.? If so, which country or countries? [SHORT TEXT RESPONSE]

**Technology Experience**

1. Does your current institution require you to use technology to teach statistics? [YES / NO / UNSURE]
2. If yes to the previous question, what kind of technology? [SHORT TEXT RESPONSE]
3. Have you ever implemented any technology in your statistical classroom? [YES / NO / UNSURE]
4. If you have, what did you implement and for what purpose? [SHORT TEXT RESPONSE]
5. If you have, rate your experience of using technology in teaching statistics from 1 to 5, where 1 = terrible and 5 = fantastic. [1 / 2 / 3 / 4 / 5 / (n/a)]
6. If you have not, what prevented you from using technology? [SHORT TEXT RESPONSE]
7. Briefly provide your opinion about using technology to teach statistics or mathematics. [SHORT TEXT RESPONSE]

**Technology Exploration**

Now you will explore some technology options for teaching introductory statistics. Please explore all of them and take as much time as you need. Feel free to bookmark or download anything you found valuable! Have fun exploring!

**Minitab**

1. Have you heard about Minitab? [YES / NO]
2. Have you used Minitab? [YES / NO]
3. Check Minitab out by visiting https://www.minitab.com/en-us/. What did you learn about Minitab from its official website? [SHORT TEXT RESPONSE]
4. You may have found out that Minitab has its own page for fully written lesson plans. If not, visit https://www.minitab.com/en-us/academic/teaching-resources. Download one or two lesson plans and check them out. What do you think about these lesson plans?

Do you think they are helpful in supporting teachers who teach statistics using Minitab? Why or why not? [SHORT TEXT RESPONSE]
5. Overall, what do you like about Minitab? [SHORT TEXT RESPONSE]
6. Overall, what don't you like about Minitab? [SHORT TEXT RESPONSE]

**Tinkerplots**

1. Have you heard about Tinkerplots? [YES / NO]
2. Have you used Tinkerplots? [YES / NO]
3. Check Tinkerplots out by visiting https://www.tinkerplots.com/. What did you learn about Tinkerplots from its official website? [SHORT TEXT RESPONSE]
4. You may have found out that Tinkerplots has its own page for class activities using Tinkerplots. If not, visit http://www.tinkerplots.com/activities. Download one or two activites and check them out. What do you think about these activities? Do you think they are helpful in supporting teachers who teach statistics using Tinkerplots? Why or why not? [SHORT TEXT RESPONSE]
5. Overall, what do you like about Tinkerplots? [SHORT TEXT RESPONSE]
6. Overall, what don't you like about Tinkerplots? [SHORT TEXT RESPONSE]

**Google Sheets**

1. Have you heard about Google Sheets? [YES / NO]
2. Have you used Google Sheets? [YES / NO]
3. Check Google Sheets out by visiting https://www.google.com/sheets/about/. What did you learn about Google Sheets from its official website? [SHORT TEXT RESPONSE]
4. You may wonder how can I teach introductory statistics using Google Sheets. Well, here is a course page created for teaching college level introductory statistics using Google Sheets. Visit http://www.comfsm.fm/~dleeling/statistics/text6.html and check it out! What do you think about the course? Do you think it is helpful in supporting teachers who teach statistics using Google Sheets? Why or why not? [SHORT TEXT RESPONSE]
5. Overall, what do you like about Google Sheets? [SHORT TEXT RESPONSE]
6. Overall, what don't you like about Google Sheets? [SHORT TEXT RESPONSE]

**Rossman/Chance Applet Collection**

1. Have you heard about any statistical web applets? [YES / NO]
2. Have you used any statistical web applets? [YES / NO]
3. If you google statistical web applets, numerous results will pop out. Today, we will visit a collection of web applets made by statistics educators. Visit http://www.rossmanchance.com/ for the main page of the applets. What did you find out by exploring the main page? [SHORT TEXT RESPONSE]
4. Visit http://www.rossmanchance.com/applets/index.html for a current list of available web applets. Check out some of them. What do you think about these applets? Do you think they are helpful in supporting teachers who teach statistics using web applets? Why or why not? [SHORT TEXT RESPONSE]
5. Overall, what do you like about this collection of web applets? [SHORT TEXT RESPONSE]

6. Overall, what don't you like about this collection of web applets? [SHORT TEXT RESPONSE]

**Rice Virtual Lab in Statistics**

1. Have you heard about any multimedia product (that combines videos, ebook, applets, and homework system as a whole package) for teaching statistics? [YES / NO]
2. Have you used any multimedia product for teaching statistics? [YES / NO]
3. Rice Virtual Lab in Statistics is one such product. Check it out by visiting http://onlinestatbook.com/rvls.html. what did you find out by exploring its official website? [SHORT TEXT RESPONSE]
4. You may have found an online textbook in introductory statistics. If not, check at http://onlinestatbook.com/index.html. What do you think about this online course page? Do you think it is helpful in supporting teachers who teach statistics? Why or why not? [SHORT TEXT RESPONSE]
5. Overall, what do you like about Rice Virtual Lab in Statistics? [SHORT TEXT RESPONSE]
6. Overall, what don't you like about Rice Virtual Lab in Statistics? [SHORT TEXT RESPONSE]

**Comparative Summary**

1. You have explored all the links in this survey. Good job! Now rank all technology options based on whether or not you'd like to integrate it into your statistics classroom. The smaller the number you give, the higher the rank. You are allowed to give the same rank to more than one option. But please adjust ranks for other items accordingly. For example, if you rank Minitab and Tinkerplots as both 1, and your next choice is Google Sheets, then you should assign a rank of 3 to Google Sheets. If you don't want to integrate any of them, rank "None" as 1 and others accordingly.

| Options | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Minitab | | | | | | |
| Tinkerplots | | | | | | |
| Google Sheets | | | | | | |
| R/C Applets | | | | | | |
| Rice Virtual Lab in Statistics | | | | | | |
| None | | | | | | |

2. Briefly explain your ranking rationale. [SHORT TEXT RESPONSE]

Appendix B

TPACK Survey

This survey is almost identical to the version used in the study. The exception is the first column (ID). In this version, all questions have been coded based on different domains in TPACK. On the actual survey, these were replaced with continuous ordering of numerical values only. At the beginning, each participant was required to input his/her study ID.

**Scoring**

Each item response is scored with a value of 1 assigned to strongly disagree, all the way to 5 for strongly agree. For each construct the participant's responses are averaged. For example, the 6 questions with TK prefixes are averaged to produce one TK Score.

**TPACK Survey**

Thank you for taking time to complete this questionnaire. Please rate each statement below to the best of your knowledge. Choose from SD (Strongly Disagree), D (Disagree), NAD (Neither Agree nor Disagree), A (Agree), and SA (Strongly Agree). Your thoughtfulness and candid responses will be greatly appreciated. Your individual name or identification number will not at any time be associated with your responses. Your responses will be kept completely confidential.

Technology is a broad concept that can mean a lot of different things. For the purpose of this questionnaire, technology is referring to

- statistical software packages such as SPSS, SAS, R, Minitab;
- educational software such as Tinkerplots and Fathom;
- spreadsheets application such as Excel and Google Sheets;
- applets (online or offline) the Rossman/Chance Applet Collection that you have explored;
- multimedia materials such as Rice Virtual Lab in Statistics that you have read about;
- computers, tablets, interactive whiteboards, or any other technology that facilitate teaching in classroom.

If you are uncertain of or neutral about your response you may always select NAD (Neither Agree or Disagree). If the item does not apply to you, you may choose NAD as well.

| ID | Statement |
|---|---|
| TK1 | I know how to solve my own technical problems. |
| TK2 | I can learn technology easily. |
| TK3 | I keep up with important new technologies. |
| TK4 | I frequently play around the technology. |
| TK5 | I know about a lot of different technologies. |
| TK6 | I have the technical skills I need to use technology. |

| ID | Statement |
|---|---|
| CK1 | I have sufficient knowledge about college level introductory statistics. |
| CK2 | I can use a statistical way of thinking. |
| CK3 | I have various ways and strategies of developing my understanding of statistics. |
| PK1 | I know how to assess student performance in class. |
| PK2 | I can adapt my teaching based upon what students currently understand or do not understand. |
| PK3 | I can adapt my teaching style to different learners. |
| PK4 | I can assess student learning in multiple ways. |
| PK5 | I can use a wide range of teaching approaches in a classroom setting. |
| PK6 | I am familiar with common student understandings and misconceptions. |
| PK7 | I know how to organize and maintain classroom management. |
| PCK1 | I can select effective teaching approaches to guide student thinking and learning in statistics. |
| TCK1 | I know about technologies that I can use for understanding and doing statistics. |
| TPK1 | I can choose technologies that enhance the teaching approaches for a lesson. |
| TPK2 | I can choose technologies that enhance students' learning for a lesson. |
| TPK3 | My teacher education program has caused me to think more deeply about how technology could influence the teaching approaches I use in my classroom. |
| TPK4 | I am thinking critically about how to use technology in my classroom. |
| TPK5 | I can adapt the use of the technologies that I am learning about to different teaching activities. |
| TPK6 | I can select technologies for my class that enhance what I teach, how I teach and what students learn. |
| TPK7 | I can use strategies that combine content, technologies and teaching approaches that I learned about in my coursework in my classroom. |
| TPK8 | I can provide leadership in helping others to coordinate the use of content, technologies and teaching approaches at my school and/or district. |
| TPK9 | I can choose technologies that enhance the content for a lesson. |
| TPACK | I can teach lessons that appropriately combine statistics, technologies and teaching approaches. |

# Appendix C

## Statistical Thinking Assessment

In this appendix, all testing items to statistical thinking assessment were listed as well as their statistical thinking, statistical reasoning, and statistical teaching. All questions, reasonings, and grading rubrics (except items 1 and 7) were adopted and modified based on the ones provided by LOCUS website. Item 7 was taken from GAISE report (Franklin et al., 2007).

During the actual interview, each participant received a version that only showed questions with some white space to write on. Participant's responses, both verbal and written, were collected by the end of the interview.

### Item 1: Warm Up

Please provide a list of topics that you normally cover in your introductory statistics course. Here are some common topics that could be covered in an introductory statistics course: Sampling and data, Descriptive statistics, Probability topics, Discrete random variables, Continuous random variables, The normal distribution, The central limit theorem, Confidence intervals, Hypothesis testing, Inference from two samples, Linear correlation and regression, The chi-square distribution, Analysis of variance. Feel free to provide more topics or subtopics if needed.

**Comments**

This item gives the researcher an idea of what the participant covers in his/her statistics class. Additionally, it is used to detect whether or not there is an imbalanced coverage between descriptive statistics and inferential statistics.

### Item 2: Statistical Question (Formulate Questions)

For their final project, students in a math class are required to answer a question by collecting data about students at their school. For which of the following questions could a random sample of students provide the best approximate answer?

(A) How many students attend the school?
(B) How many hours does each class at the school meet per year?
(C) How many text messages do students at the school send per week?
(D) Do students at this school have higher test scores than students in other schools in the district?

**Statistical Thinking**

- Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers.
- Understand that statistics can be used to gain information about a population by examining a sample of the population; generalizations about a population from a sample are valid only if the sample is representative of that population. Understand that random sampling tends to produce representative samples and support valid inferences.

**Statistical Reasoning**

At the intermediate level of the GAISE framework (Franklin et al., 2007), teachers should introduce students to the idea of a statistical question as one that anticipates variability in the data. Teachers should also introduce various methods that could be used to sample from a population. Since the data of interest will have variability, it is necessary to use random sampling to ensure the sample is representative of the population. This question assesses participants' understanding of both of these concepts.

The participant is required to select the question whose answer could be best approximated using a random sample of students at a school. The answer to option A is a deterministic one; the number of students that attend the school is a fixed number that does not vary at the time the question is being asked. It is also unclear as to how one might use random sampling to answer the question. Option B provides a question that could be answered using random sampling, but the population of interest is all the classes at the school. Therefore, it would make sense to take a random sample of classes at the school instead of students. The final distractor, option D, is a valid statistical question; however, it cannot be answered using a random sample of students from only one school.

**Statistical Teaching**

1. Why do you select this option? What's wrong with other options?
2. When you teach introductory statistics, do you mention the concept of statistical questions? How do you introduce the concept of statistical questions? For instance, do you tell students the difference between statistical questions and survey questions? If so, how do you explain the difference?
3. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Item 3: Random Sampling (Collect Data)**

A student wants to estimate the mean number of books that have been read by all students at his school over the summer. On Monday morning, he will survey the first 35 students who enter the library. Is this the best way to select a sample for this purpose?

(A) No. The student should survey students entering the library on more than one day of the week.
(B) No. The student should take a random sample of students entering the library instead.
(C) No. The student should take a random sample of students from all students, not just those entering the library.
(D) Yes. Selecting a sample in this way will not introduce the possibility of bias.

**Statistical Thinking**

- Understand that statistics can be used to gain information about a population by examining a sample of the population; generalizations about a population from a sample are valid only if the sample is representative of that population. Understand that random sampling tends to produce representative samples and support valid inferences.
- Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

**Statistical Reasoning**

In order to estimate the mean number of books that have been read by all students at the school over the summer, a student decides to survey the first 35 students who enter the library. This sampling method is known as convenience sampling and will introduce bias in the survey results because the first 35 students are not randomly selected from the population of interest. To generalize a population using a sample from that population, the sample must represent that population, and random sampling tends to produce more representative samples than convenience sampling. On the other hand, students entering the library on a Monday morning may not be representative of the larger student body since those students may be more avid readers. Surveying only those students could lead to conclusions that that might overestimate the mean number of books read by students at the school. Therefore, option C is the best answer since it proposes taking a random sample from all students, not just those entering the library.

Surveying students entering the library on more than one day of the week, as suggested in option A, does give more students the chance of being selected; however, in order for the sample to be representative of the population, all of the students must have an equal chance of being selected. Thus, Option A is incorrect, as is Option B. Option B introduces a random element, but it still excludes the students that might not enter the library. Option D is also not the correct answer for the reason discussed earlier about convenience sampling.

**Statistical Teaching**

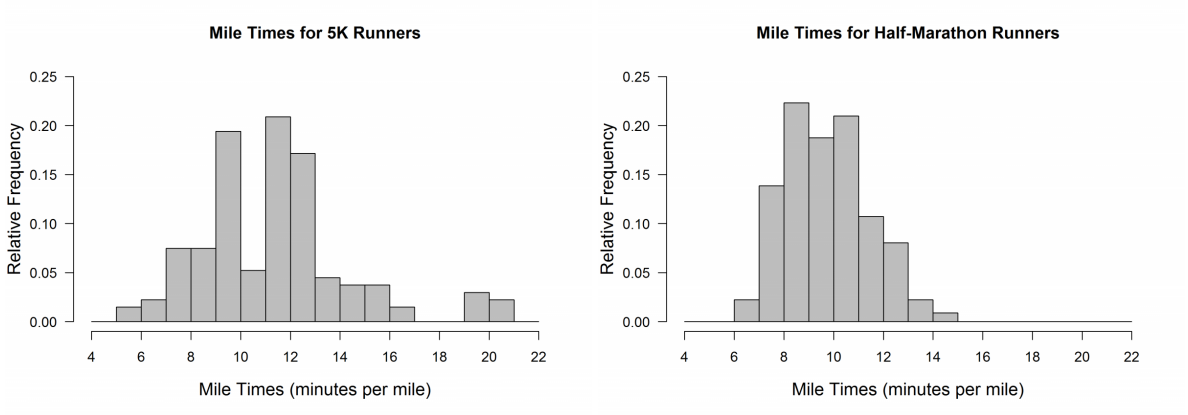
1. Why do you select this option? What's wrong with other options?
2. When you teach introductory statistics, how do you explain the concept of random sampling to students?
3. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Item 4: Center, Variability and Outliers (Analyze Data & Interpret Results)**

The city of Gainesville hosted two races last year on New Year's Day. Individual runners chose to run either a 5K (3.1 miles) or a half-marathon (13.1 miles). One hundred thirty four people ran in the 5K, and 224 people ran the half-marathon. The mile time, which is the average amount of time it takes a runner to run a mile, was calculated for each runner by dividing the time it took the runner to finish the race by the length of the race. The histograms (see Figure C.1) show the distributions of mile times (in minutes per mile) for the runners in the two races.

1. Jaron predicted that the mile times of runners in the 5K race would be more consistent than the mile times of runners in the half-marathon. Do these data support Jaron's statement? Explain why or why not.
2. Sierra predicted that, on average, the mile time for runners of the half-marathon would be greater than the mile time for runners of the 5K race. Do these data support Sierra's statement? Explain why or why not.
3. Recall that individual runners chose to run only one of the two races. Based on these data, is it reasonable to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K? Explain why or why not.

Figure C.1. *Item 4: Center, Variability and Outliers*




**Statistical Thinking**

- Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.
- Summarize numerical data sets in relation to their context, such as by:
    1. Reporting the number of observations.
    2. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement.
    3. Giving quantitative measures of center (median or mean) and variability (interquartile range or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.
    4. Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered.
- Use measures of center and measures of variability for numerical data from random samples to draw informal comparative inferences about two populations.
- Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

**Statistical Reasoning**

*Part 1*

Part 1 asks the participant to recognize that "more consistent" mile times mean that there will be less variability and then to determine which of the two given histograms shows less variability in mile times. An ideal response to part 1 is one that recognizes that there is more variability in the mile times for the 5K race than for the half-marathon and indicates that Jaron's prediction that the mile times of runners in the 5K race would be more consistent than the times for runners in the half-marathon is not correct.

Responses that correctly indicate that Jaron's prediction is incorrect but do not support the conclusion by comparing variability in the two distributions are considered to be partially correct for part 1. Also considered to be only partially correct are responses that correctly compare the variability in the two distributions but do not clearly indicate that Jaron's

prediction is not supported. Responses that indicate that Jaron's statement is supported are considered incorrect for part 1.

*Part 2*

An ideal response to part 2 is one that recognizes that the mean mile time is greater for the mile times for the 5K race than for the half-marathon and indicates that Sierra's prediction that the mile time, on average, of runners in the half-marathon would be greater than the time, on average, for runners in the 5K race is not correct.

Responses that correctly indicate that Sierra's prediction is incorrect but do not support the conclusion by comparing the centers of the two distributions are considered to be partially correct for part 2. Also considered partially correct for part 2 are responses that say that Sierra's prediction is supported and base that conclusion on the fact that the difference in where the two distributions are centered is not very large. Responses that indicate that Sierra's prediction is not supported but which do not provide an explanation are considered incorrect for part 2.

*Part 3*

In part 3, students are asked if it is reasonable to conclude that the mile time would be less if a person ran a half-marathon than if that person ran a 5K race. An ideal response indicates that it is not reasonable and provides an explanation that is based on the fact that the runners in these races chose the race in which they would run.

Responses that indicate that the stated conclusion is not reasonable but that do not explicitly mention choice of race in the explanation are considered partially correct for part 3. Responses with explanations that are not based on the study design (the way in which the data were collected) are considered incorrect for part 3.

**Statistical Teaching**

1. What are your rationales behind your responses?
2. If you are showing these two histograms along with the context to your students, what statistical concepts would you address in this example? How would you teach these concepts using this example?
3. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Item 5: Variability of Categorical Data (Analyze Data)**

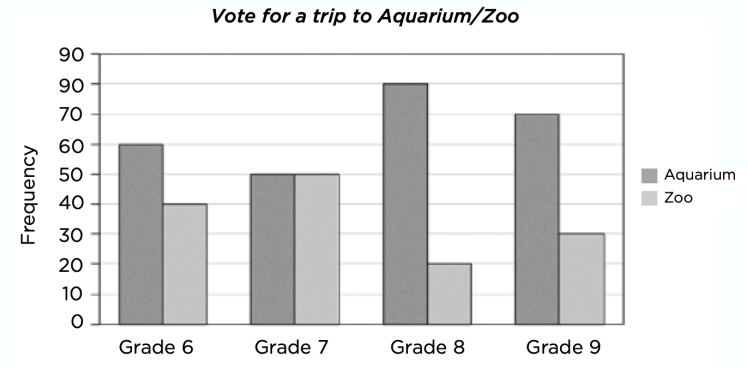
A school is planning a field trip to the aquarium or to the zoo for students in grades 6 through 9. There are 100 students in each grade level and every student was asked which place he or she would prefer to visit. The bar graphs for the four grade levels are constructed (see Figure C.2). In which grade level were the responses most consistent?

(A) Grade 6 (B) Grade 7 (C) Grade 8 (D) Grade 9

**Statistical Thinking**

- Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.
- Understand how to identify the consistency of categorical data using a bar graph.

Figure C.2. *Item 5: Variability*




**Statistical Reasoning**

The correct answer to this item is option C. The item asks participants to identify the grade level for which the responses to the survey question were the most consistent. A high level of consistency means that a large number of the students responded to the survey question the same way. In Grade 8, 80 out of 100 students chose the aquarium over the zoo, which is a higher percentage of students than selected either of the two trips in the other grade levels. Thus, Grade 8 had the most consistent responses because a relatively large number of the students chose the same trip, while the other grades were more divided in their choices.

This item tests the ability to read bar graphs and compare the consistency of responses in different groups. According to LOCUS website, almost 80% of the respondents selected option B. In Grade 7, exactly half of the survey respondents chose the zoo, and half chose the aquarium. Therefore, the heights of the bars are exactly the same in the bar graph. While the heights of the bars are consistent, this grade level represents the least consistent responses from the survey since the students are divided evenly, and there is no consensus among seventh-graders about which trip to take.

For numerical data, the spread of a distribution can be summarized by measures of variability like the interquartile range, mean absolute deviation, or standard deviation. These measures are not appropriate for categorical data, but it is still useful to describe the variability of categorical data in terms of consistency or variability, as illustrated in this item: In Grade 8, the responses were the most consistent; in Grade 7 the responses were the most variable.

**Statistical Teaching**

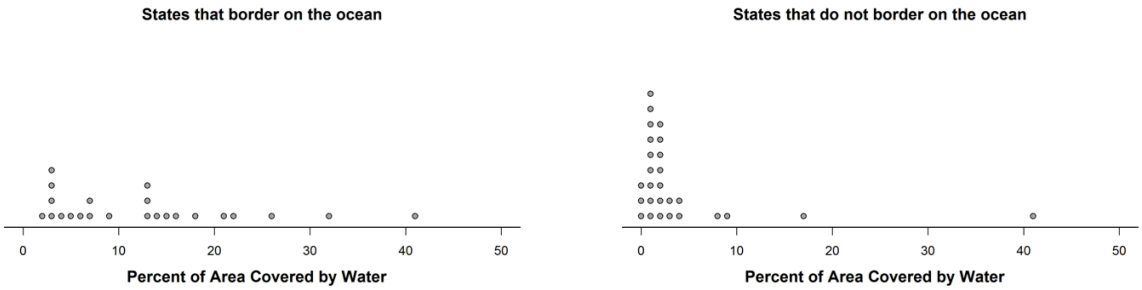
1. Why do you select this option? What's wrong with other options?
2. In your teaching, do you ever address the comparison of consistency in categorical data to students? If so, how?
3. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Item 6: Resistant Measurements to Outliers (Analyze Data)**

Carlton found data on the percent of area that is covered by water for each of the 50 states in the U.S. He made two dotplots (below)see Figure C.3) to compare the distributions for states that border an ocean and states that do not border an ocean. Which of the following is

the best statistical reason for using the median and interquartile range (IQR), rather than the mean and standard deviation, to compare the centers and spreads of these distributions?

Figure C.3. *Item 6: Outlier-Resistant Measurements*




(A) The mean and standard deviation are more strongly influenced by outliers than the median and IQR.
(B) The median and IQR are easier to calculate than the mean and standard deviation.
(C) The two groups contain different numbers of states, so the standard deviation is not appropriate.
(D) The two distributions have the same shape.

**Statistical Thinking**

- Represent data with plots on the real number line (dot plots, histograms, and box plots).
- Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.
- Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

**Statistical Reasoning**

This question requires participants to identify an explanation of why the median and interquartile range (IQR) are better for comparing the two distributions than the mean and standard deviation; implicit in this task is that participants can understand what the dotplots display. Extreme points (possible outliers) are present in both distributions, and such points have a strong influence on the mean and the standard deviation. The correct answer is Option A because the median and IQR are less sensitive to outliers.

The median and IQR may be easier to calculate than the mean and standard deviation, but this is not a statistical reason to prefer them, thus eliminating Option B as the correct answer. In professional statistical practice, all such statistics will be calculated by computer software which renders the calculations equally simple. Option C reflects a misconception that the standard deviation is only appropriate for comparing groups with equal numbers of observations, which is incorrect. Lastly, Option D is not the correct answer because 1) the distributions do not have the same shape, and 2) having the same shape is neither a requirement of nor an impediment to the use of the mean and standard deviation.

**Statistical Teaching**

1. Why do you select this option? What's wrong with other options?
2. In your teaching, do you ever address the effects of outliers on descriptive statistics? If so, what do you address and how?
3. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Item 7: Probability (Interpret Results)**

Question 1: Assume a coin is fair. If we toss the coin five times, how many heads will we get?

Question 2: You pick up a coin. Is this a fair coin?

1. Provide answers to each question.
2. How are these two questions similar?
3. How are these two questions different?
4. Is there any relationship between the concepts underlying these two problems? If so, what is it?

**Statistical Thinking**

- Understand that probability is a tool for statistics.
- Understand that statistics depends on context, variability, and uncertainty.

**Statistical Reasoning**

"Probability is an important part of any mathematical education. It is a part of mathematics that enriches the subject as a whole by its interactions with other uses of mathematics. Probability is an essential tool in applied mathematics and mathematical modeling. It is also an essential tool in statistics. The use of probability as a mathematical model and the use of probability as a tool in statistics employ not only different approaches, but also different kinds of reasoning." (Franklin et al., 2007, p. 8) This difference can be illustrated by these two problems.

"Problem 1 is a mathematical probability problem. Problem 2 is a statistics problem that can use the mathematical probability model determined in Problem 1 as a tool to seek a solution. The answer to neither question is deterministic. Coin tossing produces random outcomes, which suggests that the answer is probabilistic. The solution to Problem 1 starts with the assumption that the coin is fair and proceeds to logically deduce the numerical probabilities for each possible number of heads: 0, 1, . . . , 5. The solution to Problem 2 starts with an unfamiliar coin; we don't know if it is fair or biased. The search for an answer is experimental–toss the coin and see what happens. Examine the resulting data to see if it looks as if it came from a fair coin or a biased coin. There are several possible approaches, including toss the coin five times and record the number of heads. Then, do it again: Toss the coin five times and record the number of heads. Repeat 100 times. Compile the frequencies of outcomes for each possible number of heads. Compare these results to the frequencies predicted by the mathematical model for a fair coin in Problem 1. If the empirical frequencies from the experiment are quite dissimilar from those predicted by the mathematical model for a fair coin and are not likely to be caused by random variation in coin tosses, then we conclude that the

coin is not fair. In this case, we induce an answer by making a general conclusion from observations of experimental results." (Franklin et al., 2007, p. 8)

**Statistical Teaching**

1. What are your rationales behind your responses?
2. Some statistics instructors choose to trim down the instruction of probability in introductory statistics courses. They claim that many probability topics are not related to the subsequent instruction of inferential statistics. What do you think? What role do you think probability plays in introductory statistics?
3. How much probability do you teach to your students and why?
4. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

### Item 8: Confidence Interval for Proportions of Two Groups (Interpret Results)

Lindsey wants to use a confidence interval to estimate the difference in the proportion of females and males at her high school who have taken an honors class. She randomly selects 50 females and 50 males from her school and asks each one if he or she has taken an honors course. Of the 50 females, 23 responded yes. Of the 50 males, 19 responded yes. A 95 percent confidence interval for the difference in the proportion of females and males at her school who have taken an honors class is $0.08 \pm 0.19$.

1. Interpret the confidence interval in the context of this study.
2. The principal at Lindsey's school is interested in the results of her study but suggests that she increase the sample sizes to 100 females and 100 males. What effect will increasing the sample sizes have on Lindsey's confidence interval?

**Statistical Thinking**

- Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.
- Understand the effects of sample size and the confidence level have on the margin of error.
- Interpret the confidence interval based on the context provided.

**Statistical Reasoning**

*Part 1*

An ideal response to part 1 uses the given estimate and margin of error to find the endpoints of the confidence interval estimate of the difference in the population proportion of females and the population proportion of males who have taken an honors course and provides an appropriate interpretation for that interval. A complete response includes three parts: a reference to the difference in proportions in context, the confidence level, and correct numerical values for the endpoints of the interval. If the response includes only two of the three parts, the response is considered partially correct.

Alternatively, the confidence interval can be used to decide whether there is convincing evidence of a difference between the population proportion of males and the population proportion of females who have taken an honors course. Taking this approach, an ideal response to part 1 notes that 0 is in the confidence interval – meaning 0 is a plausible value for

the difference of proportions – and concludes that there is no convincing evidence of a difference between the population proportion of males and the population proportion of females who have taken an honors course. If the conclusion is drawn without indicating that 0 is in the confidence interval, the response is considered partially correct.

*Part 2*

An ideal response to part 2 recognizes that increasing the sample size decreases the margin of error, making the confidence interval narrower. This understanding can be demonstrated by indicating that increasing the sample size will decrease the margin of error, improve the precision of our interval, or make the interval narrower. If the response indicates that the increase in sample size decreases variability without a clear connection to the effect on the confidence interval, the response is considered only partially correct. Generic comments indicating that a larger sample size improves the study without explicitly referencing variability or the effect on the confidence interval are considered incorrect.

**Statistical Teaching**

1. What are your rationales behind your responses?
2. Some students have difficulty understanding confidence level. In your teaching, how do you approach this difficult topic? For instance, how do you tell your students what 95% confidence level means?
3. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

### Item 9: Hypothesis Testing for Means of Two Groups (Interpret Results)

A farmer conducted an experiment to find out whether a new type of fertilizer would increase the size of tomatoes grown on his farm. The farmer randomly assigned 10 tomato plants to receive the new fertilizer and 10 tomato plants to receive the old fertilizer. All other growing conditions were the same for the 20 plants. At the end of the experiment, the mean weight of tomatoes grown with the new fertilizer was 0.4 ounce heavier than the mean weight of the tomatoes grown with the old fertilizer.

1. Describe one method that the farmer could have used to randomly assign the 20 plants into groups of 10 each.
2. Based on the results, the farmer is convinced that the new fertilizer produces heavier tomatoes on average. Briefly explain to the farmer why simply comparing the two means is not enough to provide convincing evidence that the new fertilizer produces heavier tomatoes.
3. To test whether the difference of 0.4 ounce is statistically significant, a statistician calculated a $p$-value of 0.31. Based on the p-value, is there convincing evidence that the new fertilizer produces heavier tomatoes than the old fertilizer on average? Explain.

**Statistical Thinking**

- Understand the process for randomly assigning experimental units to treatments in an experiment.
- Take sampling variability into account when drawing conclusions based on data.
- Interpret the $p$-value based on the context provided.

**Statistical Reasoning**

*Part 1*

An ideal response to part 1 describes a way of assigning the 20 plants to the two fertilizers using some form of random assignment. To be considered essentially correct, the response needs to identify how the random assignment would be carried out, and the method described would need to result in two groups with 10 plants in each group. Responses that are equivalent to pulling numbers from a box or hat need to specifically mention mixing in order to be considered essentially correct.

Because the question specified groups of equal size, responses that describe methods that use random assignment but that might result in groups of different sizes (for example, flipping a coin for each plant to determine which fertilizer the plant would receive) are considered to be partially correct for part 1. Responses that do not indicate a method of random assignment (for example, just saying "randomly pick 10 plants for the first fertilizer") but do describe a method that ensures that there are 10 plants in each fertilizer group are also considered partially correct for part 1.

*Part 2*

Part 2 asks participants to explain why it is not appropriate to reach a decision based solely on the fact that one fertilizer group mean is greater than the other fertilizer group mean. An ideal response to part 2 recognizes that even if all plants received the same fertilizer, there would still be variability in tomato weights from one plant to another, and there is a need to determine if a difference of 0.4 ounce might be something that could be observed just by chance when there is no difference in the effect of the two fertilizers. To be considered essentially correct for part 2, the response must: 1) refer to sampling variability or the variability introduced by random assignment of plants to fertilizers, and 2) indicate that the observed difference in averages might be due to chance alone (the random assignment of plants to fertilizer groups). Responses that only include one of these two required elements are considered to be partially correct for part 2. A response that does not include either of these two required elements (for example, one that just says "you need to do a test"), is considered to be incorrect for part 2.

*Part 3*

Part 3 asks participants to reach a conclusion based on a given $p$-value. An ideal response to part 3 is one that correctly interprets the given $p$-value of 0.31 as large and indicates that this means that there is no convincing evidence that the new fertilizer produces heavier tomatoes on average (that the observed difference of 0.4 ounce might be due to chance alone).

Responses that indicate that there is no convincing evidence of a difference in means but that do not clearly link this conclusion to the fact that the given $p$-value is large or that base the conclusion on an incorrect interpretation of the $p$-value are considered to be partially correct for part 3.

Responses that reach an incorrect conclusion, indicating that there is convincing evidence of a difference, are incorrect for part 3. Responses that state a conclusion with no

supporting explanation or with an explanation that does not refer to the $p$-value (such as an explanation of "0.4 is small") are also considered to be incorrect for part 3.

**Statistical Teaching**

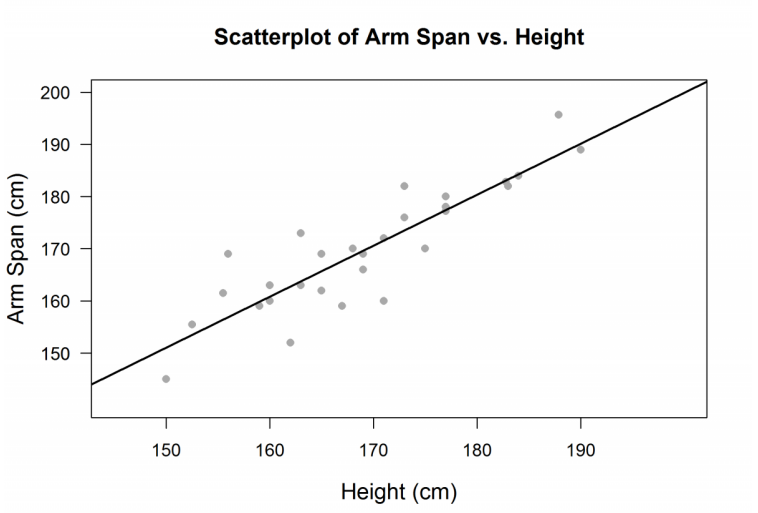
1. What are your rationales behind your responses?
2. Some students have difficulty understanding $p$-values. In your teaching, how do you approach this difficult topic? For instance, do you ever address the conceptual understanding of $p$-values? If so, what do you address and how?
3. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

**Item 10: Linear Correlation and Regression (Interpret Results)**

The heights (in centimeters) and arm spans (in centimeters) of 31 students were measured. The association between $x$ (height) and $y$ (arm span) is shown in the scatterplot (see Figure C.4). The equation of the least-squares regression line for this association is given as following:

$$\hat{y} = 4.5 + 0.977x.$$

Figure C.4. *Item 10: Height Versus Arm Span*




1. If Mike is 5 cm taller than George, what is the expected difference in their arm spans? Show your work.
2. Jane is 158 cm tall and has an arm span of 154 cm. Rhonda is 163 cm tall and has an arm span of 165 cm. Does the least-squares regression line give a more accurate predicted value for Jane or Rhonda? Explain.
3. Doug is 210 cm tall. Would you use this least-squares regression line to predict his arm span? Explain.

**Statistical Thinking**

- Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.
- Represent data on two quantitative variables on a scatter plot, and describe how the variables are related: fit a function to the data; use functions fitted to data to solve problems in the context of the data; informally assess the fit of a function by plotting and analyzing residuals; fit a linear function for a scatter plot that suggests a linear association.
- Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**Statistical Reasoning**

*Part 1*

An ideal response to part 1 recognizes that the slope of the least-squares regression line can be interpreted as the expected change in arm span associated with a 1 cm increase in height. Therefore, one only needs to multiply the given slope by 5 to obtain the expected difference in arm span for two people whose height differed by 5 cm. Responses that take this approach and that provide an explanation or include supporting work are scored as essentially correct. Another approach is to pick heights for Mike and George and plug them into the regression equation to find the expected difference in arm spans.

*Part 2*

An ideal response to part 2 indicates that the prediction of Rhonda's arm span is more accurate than the prediction of Jane's arm span and provides justification for this choice. There are two ways that one could provide a correct justification. One possible justification is based on calculating predicted values and residuals and then noting that the absolute value of the residual for Rhonda is less than the residual for Jane, indicating that the predicted arm span is closer to the actual arm span for Rhonda. A second approach that could be used to support the choice of Rhonda in part 2 uses the given scatterplot and least-squares line. One can use the information on height and arm span for Rhonda and Jane to plot points on the scatterplot. It can be noted that the point that corresponded to Rhonda's height and arm span is closer to the least-squares line than the point that corresponded to Jane's height and arm span. Because predicted arm spans are points on the least-squares line, this means that the predicted arm span would be closer to the actual arm span for Rhonda.

*Part 3*

An ideal response to part 3 recognizes that 210 cm is quite a bit greater than the height of the tallest person in the group of 31 students that were used to develop the equation of the least-squares line. Because this represents an extrapolation beyond the range of the data, the least-squares regression line should not be used to predict Doug's arm span.

**Statistical Teaching**

1. What are your rationales behind your responses?
2. Some students find the hand computation of the linear correlation coefficient $r$ and regression line tedious and impractical in real-life (since many applets or spreadsheets

can complete the calculation within seconds). What's your opinion on calculating $r$ and regression line by hand in a modern introductory statistics class?

3. Is the question above similar to what you will normally assign to your students? If yes, could you give an example of a similar problem you have assigned? If no, could you give an example of a typical problem you would assign on this topic and illustrate the differences?
4. Do you think technology integration will facilitate your teaching of this topic in any way? If so, how?

# Appendix D

## Statistical Response Accuracy

In this appendix, all participants' responses to nine statistical problems (item 2 to item 10 from the interview handout) were coded and graded as right or wrong based on the LOTUS's grading rubric. Each part of a problem was graded individually. If a response was correct, partially correct, or incorrect, it received 1, 0.5, and 0 points, respectively. In item 7, parts 2 to 4 were not graded because these were open-ended questions. The maximum number of points a participant could receive was 17. The tables below were ordered by the accuracy percentage. The last row of the tables also included the percentage of entirely correct responses to each problem.

Table D.1. *Statistical Response Accuracy (Items 2-7)*

| | | | Item 2 | | Item 3 | | Item 4 | | | | Item 5 | | Item 6 | | Item 7 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | % | Total | Key | Score | Key | Score | Part 1 | Part 2 | Part 3 | Score | Key | Score | Key | Score | Q1 | Q2 | Score |
| HH | 97.1% | 16.5 | C | 1 | C | 1 | No | No | No | 3 | C | 1 | A | 1 | 0,1,2,3,4,5 (2,3) | $z$-test | 2 |
| OZ | 91.2% | 15.5 | C | 1 | C | 1 | No | No | No | 3 | C | 1 | A | 1 | 0,1,2,3,4,5 | empirical prob | 2 |
| RM | 88.2% | 15 | C | 1 | C | 1 | No | No | No | 3 | C | 1 | A | 1 | 2,3 | empirical prob | 1.5 |
| KU | 88.2% | 15 | C | 1 | C | 1 | No | No | Yes | 2 | C | 1 | A | 1 | 0,1,2,3,4,5 (2,3) | $z$-test | 2 |
| DE | 85.3% | 14.5 | C | 1 | C | 1 | No | No | No | 3 | B | 0 | A | 1 | 0,1,2,3,4,5 (2,3) | empirical prob | 2 |
| WG | 76.5% | 13 | BC | 0.5 | C | 1 | No | No | Yes | 2 | B | 0 | A | 1 | 0,1,2,3,4,5 (2,3) | empirical prob | 2 |
| AQ | 76.5% | 13 | C | 1 | C | 1 | No | No | Yes | 2 | C | 1 | A | 1 | 2,3 | empirical prob | 1.5 |
| PK | 73.5% | 12.5 | C | 1 | C | 1 | No | No | Yes | 2 | B | 0 | A | 1 | 2,3 | $z$-test | 1.5 |
| PV | 73.5% | 12.5 | C | 1 | C | 1 | No | No | Yes | 2 | B | 0 | A | 1 | 0,1,2,3,4,5 | empirical prob | 2 |
| FX | 73.5% | 12.5 | C | 1 | C | 1 | No | No | Yes | 2 | B | 0 | A | 1 | 0,1,2,3,4,5 (2,3) | empirical prob | 2 |
| MP | 70.6% | 12 | CD | 0.5 | C | 1 | No | No | Yes | 2 | B | 0 | A | 1 | 0,1,2,3,4,5 | experiment | 1.5 |
| RY | 67.6% | 11.5 | C | 1 | C | 1 | No | No | Yes | 2 | C | 1 | A | 1 | 2,3 | unknown | 1 |
| QK | 64.7% | 11 | C | 1 | C | 1 | No | No | Yes | 2 | B | 0 | A | 1 | 2,3 | empirical prob | 1.5 |
| YW | 52.9% | 9 | D | 0 | B | 0 | No | No | Yes | 2 | B | 0 | A | 1 | 2.5 | unknown | 1 |
| GG | 38.2% | 6.5 | BCD | 0.5 | C | 1 | No | No | Yes | 2 | B | 0 | B | 0 | a lot | experiment | 0.5 |
| | | | 73.3% | | 93.3% | | 100% | 100% | 26.7% | | 40.0% | | 93.3% | | 53.3% | 73.3% | |

Table D.2. *Statistical Response Accuracy (Items 8-10)*

| | | | Item 8 | | | Item 9 | | | | Item 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | % | Total | Part 1 | Part 2 | Score | Part 1 | Part 2 | Part 3 | Score | Part 1 | Part 2 | Part 3 | Score |
| HH | 97.1% | 16.5 | cont CI | narrower | 1.5 | Excel sampling | variability+chance+CI | No | 3 | slope, 4.885 | visual, rhonda | No | 3 |
| OZ | 91.2% | 15.5 | cont CI | narrower | 1.5 | head vs tail | variability+chance | No | 2.5 | slope, 4.885 | visual, no result | No | 2.5 |
| RM | 88.2% | 15 | nu/co CI+conf lvl | narrower | 1.5 | random # | chance | No | 2 | (.977)(5) = 4.885 | algebra, rhonda | No | 3 |
| KU | 88.2% | 15 | nume+cont CI | narrower | 1.5 | blind draw | variability+chance | No | 2.5 | scaling, 4.885 | estimate, rhonda | No | 3 |
| DE | 85.3% | 14.5 | cont CI | narrower | 1.5 | random # | variability | No | 2 | slope, (.977)(5)≈5 | visual, rhonda | No | 3 |
| WG | 76.5% | 13 | nu/co CI+conf lvl | narrower | 1.5 | random # | prac insignificant | No | 2 | scaling, 4.885 | algebra, rhonda | No | 3 |
| AQ | 76.5% | 13 | nume CI | narrower | 1.5 | random # | variability+CI/$p$-value | No | 2 | algebra, no result | algebra, no result | No | 2 |
| PK | 73.5% | 12.5 | cont CI | narrower | 1.5 | blind draw | chance | No | 2.5 | algebra, 4.885 | algebra, rhonda | Yes | 2 |
| PV | 73.5% | 12.5 | cont CI | narrower | 1.5 | blind draw | variability+chance | No | 2.5 | algebra, no result | algebra, rhonda | Yes | 1.5 |
| FX | 73.5% | 12.5 | cont CI | narrower | 1.5 | blind draw | variability+prac insig | No | 2 | algebra, no result | algebra, no result | No | 2 |
| RY | 67.6% | 11.5 | cont CI | narrower | 1.5 | odd vs even | prac insignificant | No | 2 | algebra, no result | algebra, no result | Yes | 1 |
| MP | 70.6% | 12 | nume+cont CI | narrower | 1.5 | every other plant | chance+var+prac insig | No | 2.5 | (.977)(5), no result | visual, no result | No | 2 |
| QK | 64.7% | 11 | nume CI | narrower | 1.5 | random # | chance | No | 2 | algebra, no result | algebra, no result | Yes | 1 |
| YW | 52.9% | 9 | nume CI | narrower | 1.5 | random selection | stat insignificant | No | 2 | need heights | algebra, no result | No | 1.5 |
| GG | 38.2% | 6.5 | nume CI | narrower | 1 | blind draw | why not convincing? | No | 1.5 | plug in $x$=5 | irrelevant | Yes | 0 |
| | | | 0.0% | 100% | | 13.3% | 33.3% | 100% | | 46.7% | 46.7% | 66.7% | |

# Appendix E

## Statistical Response Thinking

In this appendix, all participants' responses to nine statistical problems during the interviews (item 2 to item 10 from the interview handout) were coded and categorized as statistical thinking, mathematical thinking, a mix of both, or neither. Each part of a problem was evaluated individually. A response that was identified as one type of thinking typically received 0.5 points. If the thinking was evident and matched with the six main themes (Problem-Solving Process, Variability, Context, Data Production, Transnumeration, and Probabilistic Thinking) of this study, the corresponding response received 1 point. The maximum number of points for statistical thinking responses and mathematical thinking responses were 17.5 and 13.5, respectively.

The first table below summarized the overall types of thinking that participating mathematics teachers revealed. The tables following presented the coded thinking in color where statistical thinking was highlighted in orange, mathematical thinking was highlighted in blue, mixed thinking was highlighted in olive green–a mix of orange and blue, and neither statistical nor mathematical thinking was not colored. A strong type of thinking was colored in high saturation. In the header of tables:

- $S$: the statistical thinking score based on coding.
- $M$: the mathematical thinking score based on coding.
- $z_s$: the normalized statistical thinking score compared to half of the total statistical thinking points where $z_s = (S - 8.75)/(\text{stdev of } S)$.

Table E.1. *Statistical Response Thinking Overview*

| ID | $ST$ | $z_s$ | $z_m$ | $S$ | $M$ |
|---|---|---|---|---|---|
| HH | 3.22 | 1.68 | -1.53 | 14 | 3.5 |
| OZ | 2.90 | 1.36 | -1.53 | 13 | 3.5 |
| DE | 2.56 | 1.20 | -1.33 | 12.5 | 4 |
| MP | 1.12 | 0.40 | -0.72 | 10 | 5.5 |
| KU | 0.87 | 0.56 | -0.31 | 10.5 | 6.5 |
| RM | 0.18 | 0.08 | -0.10 | 9 | 7 |
| PV | -0.18 | -0.08 | 0.10 | 8.5 | 7.5 |
| RY | -0.66 | -0.56 | 0.10 | 7 | 7.5 |
| FX | -0.75 | -0.24 | 0.51 | 8 | 8.5 |
| WG | -0.80 | -0.08 | 0.72 | 8.5 | 9 |
| AQ | -0.87 | -0.56 | 0.31 | 7 | 8 |
| QK | -2.01 | -0.88 | 1.13 | 6 | 10 |
| YW | -2.08 | -1.36 | 0.72 | 4.5 | 9 |
| GG | -2.65 | -1.52 | 1.13 | 4 | 10 |
| PK | -3.10 | -1.36 | 1.74 | 4.5 | 11.5 |

- $z_m$: the normalized mathematical thinking score compared to half of the total mathematical thinking points where $z_m = (M - 7.25)/(\text{stdev of } M)$.
- $ST$: the primary thinking where $ST = z_s - z_m$.

All tables were ordered by $ST$ in descending order. A positive $ST$ indicated a tendency of statistical thinking, whereas a negative $ST$ implied mathematical thinking. A $ST$ around 0 suggested either a combination of or a lack of both types of thinkings.

Table E.2. *Statistical Response Thinking (Items 2-5)*

| | Item 2 | | | Item 3 | | | Item 4 | | | | | Item 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Key | $S$ | $M$ | Key | $S$ | $M$ | Part 1 | Part 2 | Part 3 | $S$ | $M$ | Key | $S$ | $M$ |
| HH | C (use CI+pop in D) | 0.5 | 0.5 | C (pop) | 0.5 | 0 | No (read histogram) | No (skewness+effect) | No (diff runners) | 3 | 0 | C (more aquarium) | 1 | 0 |
| OZ | C (deter in AB+pop in D) | 0.5 | 0.5 | C (pop) | 0.5 | 0 | No (read histogram) | No (skewness+effect) | No (diff runners) | 3 | 0 | C (more aquarium) | 1 | 0 |
| DE | C (variability+pop in ABCD) | 1 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (skewness) | No (diff runners) | 2.5 | 0 | B (same heights) | 0 | 1 |
| MP | CD (data+deterministic in AB) | 0.5 | 0.5 | C (pop) | 0.5 | 0 | No (read histogram) | No (average+skewness) | Yes (by Part 2) | 1.5 | 1.5 | B (same heights) | 0 | 1 |
| KU | C (pop in D) | 0.5 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (mean) | Yes (by Part 2) | 1.5 | 1 | C (more aquarium) | 1 | 0 |
| RM | C (pop in ABCD) | 1 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (average) | No (diff runners) | 2 | 0.5 | C (more aquarium) | 1 | 0 |
| PV | C (pop in D) | 0.5 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (average) | Yes (by Part 2) | 1 | 1.5 | B (same heights) | 0 | 1 |
| RY | C (pop in D) | 0.5 | 0 | C (mini pop) | 1 | 0 | No (read histogram) | No (average+skewness) | Yes (by Part 2) | 1.5 | 1.5 | C (more aquarium) | 1 | 0 |
| FX | C (pop in D) | 0.5 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (average) | Yes (by Part 2) | 1 | 1.5 | B (same heights) | 0 | 1 |
| WG | BC (use CI+pop in D) | 0.5 | 0.5 | C (pop) | 0.5 | 0 | No (read histogram) | No (average) | Yes (by Part 2) | 1 | 1.5 | B (same heights) | 0 | 1 |
| AQ | C | 0 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (average + skewness) | Yes (by Part 2) | 1.5 | 1.5 | C (more aquarium) | 1 | 0 |
| QK | C (variability+pop in D) | 1 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (average) | Yes (by Part 2) | 1 | 1.5 | B (same heights) | 0 | 1 |
| YW | D (deterministic in B) | 0 | 0.5 | B (subset) | 0 | 1 | No (read histogram) | No (average+skewness) | Yes (by Part 2) | 1.5 | 1.5 | B (same heights) | 0 | 1 |
| GG | BCD | 0 | 0 | C (pop) | 0.5 | 0 | No (read histogram) | No (skewness) | Yes (by Part 2) | 1.5 | 1 | B (same heights) | 0 | 1 |
| PK | C (use CI+pop in D) | 0.5 | 0.5 | C (pop) | 0.5 | 0 | No (read histogram) | No (average) | Yes (by Part 2) | 1 | 1.5 | B (same heights) | 0 | 1 |

Table E.3. *Statistical Response Thinking (Items 6-7)*

| | Item 6 | | | Item 7 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | Key | $S$ | $M$ | Part 1 Q1 | Part 1 Q2 | Part 2 | Part 3 | Part 4 | $S$ | $M$ |
| HH | A (skewness+stdev formula) | 0.5 | 0.5 | 0,1,2,3,4.5 (2,3) | $z$-test (0.5?) | fairness, $P(H)$ | reversed "fairness" | probability | 2.5 | 1 |
| OZ | A (skewness+effect+stdev formula) | 0.5 | 0.5 | 0,1,2,3,4.5 | empirical probability (0.5?) | fairness | reversed "fairness" | probability | 2 | 1 |
| DE | A (skewness+stdev formula) | 0.5 | 0.5 | 0,1,2,3,4.5 (2,3) | empirical probability | fairness, binomial | reversed "fairness", data | probability | 2.5 | 1 |
| MP | A (skewness+effect) | 1 | 0 | 0,1,2,3,4.5 | experiment | fairness, coin, $P(H)$ | theoretical vs empirical | probability | 2 | 1 |
| KU | A (skewness+stdev formula) | 0.5 | 0.5 | 0,1,2,3,4.5 (2,3) | $z$-test (0.5?) | fairness, coin, $P(H)$, indeter | reversed "fairness", descr vs infer | probability, descr & infer | 3.5 | 1 |
| RM | A (skewness+stdev formula) | 0.5 | 0.5 | 2,3 | empirical probability (0.5?) | fairness | reversed "fairness" | probability | 0.5 | 2.5 |
| PV | A (skewness) | 0.5 | 0 | 0,1,2,3,4.5 | empirical probability (0.5?) | fairness, $P(H)$ | Q2 promotes statistical thinking | probability | 2.5 | 1 |
| RY | A (skewness+MAD) | 0.5 | 0.5 | 2,3 | unknown | coin | reversed "fairness" | probability | 0.5 | 1 |
| FX | A (skewness+stdev formula) | 0.5 | 0.5 | 0,1,2,3,4.5 (2,3) | empirical probability (0.5?) | fairness | reversed "fairness" | probability | 2.5 | 1 |
| WG | A (stdev formula) | 0 | 0.5 | 0,1,2,3,4.5 (2,3) | empirical probability (0.5?) | fairness, coin, independence | reversed "fairness" | probability | 3 | 1 |
| AQ | A (skewness) | 0.5 | 0 | 2,3 | experiment | indeterministic | #heads, fairness | probability | 1 | 1.5 |
| QK | A | 0 | 0 | 2,3 | empirical probability (0.5?) | coin, $P(H)$ | theoretical vs empirical | probability | 1 | 2.5 |
| YW | A (skewness) | 0.5 | 0 | 2.5 | unknown | coin | reversed "fairness" | unsure | 0 | 1 |
| GG | B | 0 | 0 | a lot | experiment | coin | theoretical vs empirical | probability | 1 | 1.5 |
| PK | A (CV) | 0 | 0.5 | 2,3 | $z$-test (0.5?) | fairness, coin | reversed "fairness" | probability | 0.5 | 2.5 |

Table E.4. *Statistical Response Thinking (Items 8-10)*

| | Item 8 | | | | Item 9 | | | | | Item 10 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Part 1 | Part 2 | $S$ | $M$ | Part 1 | Part 2 | Part 3 | $S$ | $M$ | Part 1 | Part 2 | Part 3 | $S$ | $M$ |
| HH | coCI | narrower (sampling error) | 1.5 | 0 | Excel sampling | var+chance (CI) | No (big $p$-value) | 1.5 | 1.5 | slope, 4.885 | residual, rhonda | No (out of range) | 3 | 0 |
| OZ | coCI | narrower (est precision) | 1 | 0 | head vs tail | var+chance | No (big $p$-value) | 1.5 | 1.5 | slope, 4.885 | residual, no result | No (out of range) | 3 | 0 |
| DE | coCI | narrower (est precision) | 1 | 0 | random# | variability | No (big $p$-value) | 1.5 | 1.5 | slope, (.977)(5)≈5 | residual, rhonda | No (out of range) | 3 | 0 |
| MP | nu/coCI | narrower (est precision) | 1 | 0 | every other plant | chance+var+prac insig | No (big $p$-value) | 1.5 | 1.5 | (.977)(5), no result | residual, no result | No (out of range) | 2 | 0 |
| KU | nu/coCI | narrower (est precision) | 1 | 0 | blind draw | var+chance | No (big $p$-value) | 1.5 | 1.5 | scaling, 4.885 | estimate, rhonda | No (yes first) | 0.5 | 2.5 |
| RM | nu/coCI+conf lvl | narrower (formula) | 1 | 1 | random# | chance | No (big $p$-value) | 1.5 | 1.5 | (.977)(5) = 4.885 | algebra, rhonda | No (out of range) | 1 | 1 |
| PV | coCI | narrower (formula) | 0.5 | 1 | blind draw | var+chance | No (depends) | 2.5 | 0.5 | algebra, no result | alg+residual, rhonda | Yes (plug in) | 0.5 | 2.5 |
| RY | coCI | narrower (est precision) | 1 | 0 | odd vs even | prac insig | No (big $p$-value) | 1 | 1.5 | algebra, no result | algebra, no result | Yes (plug in) | 0 | 3 |
| FX | coCI | narrower (formula) | 0.5 | 1 | blind draw | var+prac insig | No (big $p$-value) | 1.5 | 1.5 | algebra, no result | algebra, no result | No (out of range) | 1 | 2 |
| WG | nu/coCI+conf lvl | narrower (formula) | 1 | 1 | random# | chance+prac insig | No (big $p$-value) | 1.5 | 1.5 | scaling, 4.885 | algebra, rhonda | No (out of range) | 1 | 2 |
| AQ | nuCI | narrower (est precision) | 0.5 | 1 | random# | var+CI/$p$-value | No (big $p$-value) | 1 | 2 | algebra, no result | algebra, no result | No (out of range) | 1 | 2 |
| QK | nuCI | narrower (est precision) | 0.5 | 1 | random# | chance | No (chance+big $p$) | 2 | 1 | algebra, no result | algebra, no result | Yes (plug in) | 0 | 3 |
| YW | nuCI | narrower (est precision) | 0.5 | 1 | random selection | pract insig | No (big $p$-value) | 1 | 1.5 | need heights | algebra, no result | No (out of range) | 1 | 1.5 |
| GG | nuCI | narrower (formula) | 0 | 2 | blind draw | why not convincing? | No (big $p$-value) | 0.5 | 2.5 | plug in $x$=5 | residual, irre work | Yes (plug in) | 0.5 | 2 |
| PK | coCI | narrower (formula) | 0.5 | 1 | blind draw | chance | No (big $p$-value) | 1.5 | 1.5 | algebra, 4.885 | algebra, rhonda | Yes (plug in) | 0 | 3 |

# Appendix F

## Online Survey Select Responses

In this appendix, some participants' responses to the two online surveys were combined and presented altogether. The table was ordered by participants' ID.

Table F.1. *Online Survey Select Results 1*

| ID | gender | age | degree | undergrad | grad | #stat | #math | #CS | #edu | language | #yrs teach | #yrs stat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AQ | Male | ≥50 | PhD | Math | Math | 1 | 30 | 2 | 0 | n/a | 25 | 10 |
| DE | Male | 30-39 | PhD | Math | Math | 1 | 30 | 2 | 1 | Chinese | 13 | 3 |
| FX | Female | ≥50 | Master | Math | Math | 0 | 20 | 0 | 0 | French | 20 | 10 |
| GG | Female | 30-39 | PhD | Math Education | Math Education | 8 | 20 | 2 | 10 | Korean | 7 | 1 |
| HH | Female | 30-39 | Master | Math (Stat Track) | Applied Math | 7 | 15 | 2 | 0 | n/a | 3 | 3 |
| KU | Female | ≥50 | Master | Math | Math | 0 | 25 | 0 | 0 | French | 20 | 10 |
| MP | Male | 30-39 | Master | Math | Math | 0 | 18 | 0 | 0 | n/a | 5 | 3 |
| OZ | Female | ≥50 | Bachelor | Math | n/a | 0 | 15 | 0 | 0 | n/a | 35 | 15 |
| PK | male | 40-49 | Master | Engineering | Engineering | 1 | 5 | 1 | 0 | n/a | 5 | 5 |
| PV | Female | 30-39 | Master | Finance | Applied Math | 1 | 23 | 0 | 10 | Russian | 4 | 2 |
| QK | Male | 40-49 | EdD | Math | Math Education | 2 | 25 | 1 | 10 | Spanish | 11 | 11 |
| RM | Female | 30-39 | PhD | Math (Stat Track) | Math/Stat Education | 10 | 10 | 1 | 5 | n/a | 8 | 3 |
| RY | Female | 30-39 | Master | Math | Applied Math | 0 | 20 | 3 | 0 | Hindi | 5 | 5 |
| WG | Male | ≥50 | PhD | Math | Math | 1 | 30 | 1 | 0 | French | 25 | 15 |
| YW | Female | 40-49 | Master | Chemistry | Applied Math | 1 | 20 | 1 | 0 | Spanish | 15 | 2 |

Table F.2. *Online Survey Select Results 2*

| ID | tech eva | Minitab | Tinkerplots | Sheets | R/C | Rice | None | TK | CK | PK | PCK | TCK | TPK | TPACK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AQ | 3 | 5 | 4 | 1 | 6 | 2 | 3 | 1.33 | 5.00 | 5.00 | 5.00 | 3.00 | 3.00 | 3.00 |
| DE | 3 | 2 | 4 | 2 | 5 | 5 | 1 | 2.67 | 4.00 | 4.00 | 4.00 | 3.00 | 3.56 | 4.00 |
| FX | n/a | 3 | 2 | 4 | 5 | 1 | 6 | 1.00 | 4.00 | 4.14 | 4.00 | 1.00 | 1.22 | 3.00 |
| GG | 5 | 1 | 4 | 2 | 3 | 5 | 6 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 |
| HH | 4 | 4 | 6 | 2 | 3 | 1 | 5 | 4.17 | 4.67 | 3.57 | 3.00 | 5.00 | 3.67 | 4.00 |
| KU | 3 | 4 | 3 | 6 | 1 | 5 | 2 | 2.00 | 4.33 | 4.29 | 4.00 | 3.00 | 3.00 | 3.00 |
| MP | n/a | 6 | 5 | 3 | 1 | 2 | 4 | 3.67 | 4.00 | 3.86 | 4.00 | 3.00 | 3.78 | 3.00 |
| OZ | n/a | 4 | 3 | 6 | 2 | 1 | 5 | 1.00 | 5.00 | 5.00 | 5.00 | 1.00 | 1.22 | 1.00 |
| PK | 3 | 4 | 3 | 5 | 2 | 1 | 6 | 2.33 | 4.00 | 3.71 | 3.00 | 3.00 | 3.67 | 3.00 |
| PV | 5 | 2 | 5 | 4 | 3 | 1 | 6 | 3.33 | 4.00 | 4.29 | 4.00 | 5.00 | 3.33 | 4.00 |
| QK | 5 | 3 | 4 | 5 | 2 | 1 | 6 | 2.83 | 4.00 | 4.86 | 4.00 | 5.00 | 3.89 | 4.00 |
| RM | 4 | 5 | 4 | 1 | 2 | 3 | 6 | 5.00 | 4.67 | 4.57 | 4.00 | 4.00 | 4.11 | 4.00 |
| RY | 3 | 4 | 3 | 6 | 2 | 1 | 5 | 1.83 | 4.00 | 3.57 | 3.00 | 3.00 | 3.00 | 3.00 |
| WG | 3 | 2 | 6 | 5 | 4 | 1 | 3 | 1.50 | 4.00 | 5.00 | 5.00 | 3.00 | 3.56 | 3.00 |
| YW | n/a | 1 | 2 | 5 | 6 | 3 | 4 | 1.00 | 4.00 | 4.00 | 3.00 | 3.00 | 1.22 | 3.00 |