Mathematics Education Journal
https://doi.org/10.22342/mej.v19i3.pp437-464

Mathematics Pre-Service Teachers' Understanding of Inferential Statistics: A Case Study in Singapore

Qiaozi Peng, Tin Lam Toh*, Ying Zhu

National Institute of Education, Nanyang Technological University, Singapore, Singapore
*Email: tinlam.toh@nie.edu.sg

Abstract

In this paper, we report our study on identifying mathematics pre-service teachers' understanding (and misconceptions) of concepts in inferential statistics through case study methodology of an entire cohort of nine beginning undergraduate students in a teacher education course. Multiple-choice questions and open-ended questions were used to elicit their responses on sampling distribution, Central Limit Theorem, and concepts related to hypothesis testing. The students' responses show their understanding of sampling distribution and Central Limit Theorem, but lack of understanding of concepts related to hypothesis testing. Their knowledge of hypothesis testing was characterized by their procedural approach to perform hypothesis testing. Some suggestions on teaching of statistics in the school mathematics curriculum are also provided.

Keywords: Basic Literacy Skills, Hypothesis Testing, Inferential Statistics, Statistical Thinking

How to Cite: Peng. Q., Toh, T. L., & Zhu, Y. (2025). Mathematics pre-service teachers' understanding of inferential statistics: A case study in Singapore. *Mathematics Education Journal*, *19*(3), 437-464. https://doi.org/10.22342/mej.v19i3.pp437-464

INTRODUCTION

In this paper, we report our study of mathematics pre-service teachers' understanding (and misunderstanding) of statistics concepts, in particular, inferential statistics. Statistics education has become increasingly important because of its emphasis on decision-making in a world that is filled with much uncertainty and variation (Stapor, 2020). Statistics plays a crucial role in the development of thoughtful citizens who can take on cooperative, responsible, and steadfast roles supported by science and a shared perspective (Da Silva et al., 2021). It is thus not surprising that statistics education encompasses a wide range of disciplines and is recognized as a fundamental component of a well-rounded education (Weiland et al., 2019). The emphasis on the concept of teaching statistics began only relatively recently compared to other branches of mathematics; it started to become visible following the first International Conference on Teaching Statistics (ICOTS) held in 1982 (Batanero & Borovcnik, 2016).

Over the past five decades of the late 20th century, the scope of statistics taught in the English-speaking school curriculum has significantly expanded to a significant portion of the mathematics taught to all 5 to 16-year-old students as well as an essential component of other academic subjects (Holmes, 2003). However, the mathematics education paradigms gave little consideration to probability and statistics, although probability and statistics has been recognized as a critical part of mathematics education due to their utility values in 20th century (Vere-Jones, 1995). In the modern world, the demands of a data-centric society have emphasized the importance of statistics education (Engel, 2017). Efforts have been made to integrate statistics into the mathematics curriculum in various countries (e.g., Hijazi & Shaqlaih, 2023; Shi et al., 2009).

Received October 4, 2024; Revised November 22, 2024; Accepted March 9, 2025

Inferential statistics has been included into the pre-university mathematics curriculum in Singapore for many decades. One of the main goals of pre-university mathematics in Singapore is to equip students with essential concepts and skills necessary for success in tertiary studies (Ministry of Education Singapore (MOE), 2020a). This study was conducted to explore the understanding (and misunderstanding) of key concepts of inferential statistics among beginning undergraduate students in Singapore, who have not yet been exposed to university-level statistics courses, as a gauge of their knowledge of pre-university statistics.

Statistics Education in Singapore

Singapore mathematics curriculum at primary and secondary levels incorporate only descriptive statistics (MOE, 2013; MOE, 2020b). At the pre-university level, mathematics curriculum includes both probability and inferential statistics, covering topics such as sampling distribution, Central Limit Theorem and hypothesis testing (MOE, 2020a). As primary and secondary students learn only descriptive statistics, they first encounter inferential statistics at the pre-university level. Note that students are not exposed to informal inference (such as intuitive ideas of sampling) in earlier school studies. They first engage directly with formal statistical inference during their pre-university education.

Our collective knowledge of the Singapore mathematics classrooms in the Singapore mathematics classrooms shows that statistics is usually taught as the final chapter of the mathematics course. Thus, teachers have to rush through statistics topics in order to prepare their students for the national examinations, probably not paying much attention to the statistical concept development. Together with the use of graphing calculators, it is common knowledge that teachers focus on procedural computation rather than conceptual understanding. The relationship between procedural knowledge and conceptual knowledge is complex, procedural computation and application of formulae do not necessarily lead to conceptual understanding (Braithwaite & Sprague, 2021; Rumsey, 2002). The inclusion of calculators was meant to focus on higher order thinking skills (MOE, 2007). However, some researchers have asserted that excessive reliance on calculators could hinder students from comprehending fundamental concepts (Naseer, 2015; Khalid & Embong, 2019).

Statistical Thinking

Although Statistics began as part of mathematics, some researchers have argued that it is distinct from mathematics (e.g., Page & Moore, 1988). New perspectives on methods of instruction and the acquisition of statistics knowledge as distinct from mathematics are essential (Groth, 2015). Many countries, such as China, Australia, the United States, New Zealand, and Israel, have undergone a paradigm shift from in school education, moving from traditional mathematics-centered approaches to a focus on understanding data and statistical thinking and literacy in teaching statistics (e.g., Hijazi & Shaqlaih, 2023; Pfannkuch & Ben-Zvi, 2011; Zhang & Stephens, 2016).

Statistics involves interpreting data, which differs from mathematical thinking. Understanding the scope and limitations of data and knowing how to ask meaningful questions about data has become extremely important (Bargagliotti et al., 2020). The pivotal role of statistical thinking has received attention as part of statistics education, e.g., Guidelines for Assessment and Instruction in Statistics Education (GAISE) framework of the United States (Bargagliotti et al., 2020; Franklin et al., 2005).

There are various interpretations of statistical thinking. Moore (1990) identified five core elements of statistical thinking: (1) The omnipresence of variation in process; (2) The need for data about processes; (3) The design of data production with variation in mind; (4) The quantification of variation; and (5) The explanation of variation.

Snee (1990) defined statistical thinking, in the context of quality improvement, as 'thought processes' which acknowledge the omnipresence of variation in our daily lives and in all our actions. Statistical thinking focuses on the significance of identifying, describing, measuring, managing, and minimizing variation as means to enhance and create opportunities for improvement. Snee succinctly summarized his definition in a schematic manner (Figure 1).

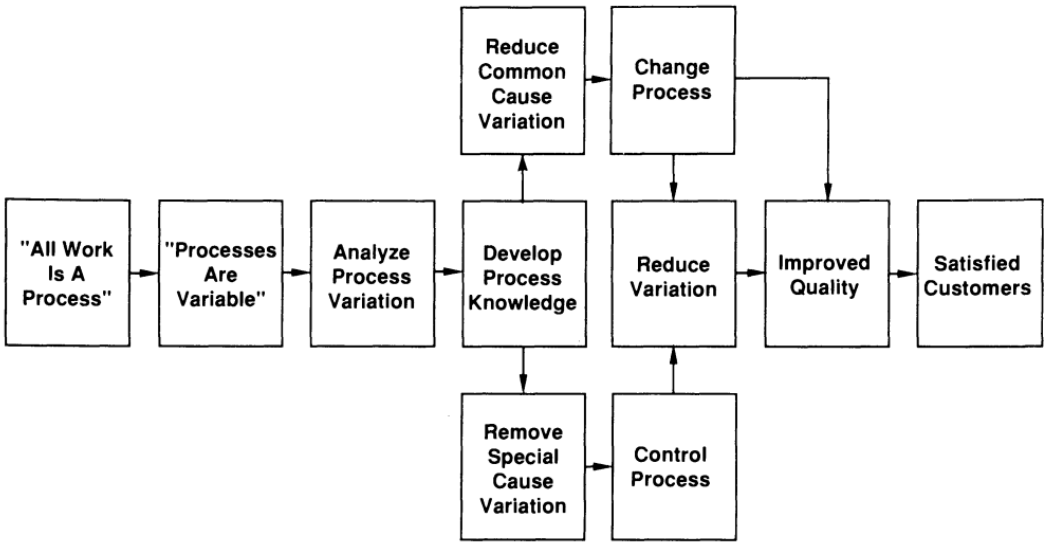

Figure 1. Statistical thinking in quality improvement (Snee, 1990, p. 118)

Garfield et al. (2003) described three components of statistical thinking: (1) Understanding the underlying principles and methods of statistical investigations, recognizing the importance of variation and knowing how to analyze data using numerical summaries and visual displays; (2) Understanding the significance of sampling and making inferences from samples to populations, using models to simulate random events and estimate probabilities and knowing when and how to use inferential tools in the investigative process; and (3) Considering the context of a problem, conducting investigations, drawing conclusions based on the problem's specific circumstances, and acknowledging and comprehending the complete process.

Chance (2002) believed that statistical thinking involves the following components: (1) the ability to view the entire process, including its iterative nature; (2) the ability to understand the

significance of variation within that process, and (3) the ability to explore data beyond standard methods and (4) generate new questions within specific contexts. These aspects make statistical thinking go beyond mere statistical literacy and reasoning.

In examining the various descriptions of statistical thinking among the various researchers, some common ideas which form the essence of statistical thinking can be identified: (1) Recognizing the omnipresence of variation and understanding the inherent variability in individuals and measurements; and understanding variation in the design of data production; (2) Appreciating the importance of empirical evidence and data examination and prioritizing the collection, analysis, and interpretation of data to derive meaningful insights; (3) Quantifying variation through mathematical descriptions of random processes using probability; (4) Recognizing the data within the context of the problem to identify systematic effects behind random variability.

One of the challenges in teaching statistics is that students often associate it with mathematics, expecting a focus on numbers, calculations, formulas, and finding a single correct answer (Ben-Zvi & Garfield, 2004; Dani & Quraan, 2023). Statistics encompasses working with complex data, exploring multiple interpretations based on diverse assumptions, and demanding proficient writing and communication abilities. Having a strong mathematics background does not guarantee proficiency in statistical thinking (Hannigan et al., 2013). Further, students proficient in computation might struggle to derive meaningful interpretations from experimental results (Delcham & Sezer, 2010), as mathematical and statistical thinking are not synonymous. Mathematical thinking prioritizes abstract patterns and structures, usually dismissing context as extraneous details (Cobb & Moore, 1997). In contrast, statistical thinking focuses on variability and the understanding that real-world data hold contextual significance. School mathematics could possibly lead students to adopt a deterministic perspective when approaching quantitative entities (Scheaffer, 2006).

Descriptive and Inferential Statistics

Descriptive statistics and inferential statistics represent two primary approaches in statistical analysis and interpretation of data (Mishra et al., 2019). As descriptive statistics enables researchers to make conclusions specifically about the data at hand, they do not extend to drawing conclusions about the population beyond the dataset (McTavish & Loether, 2018).

The modern application of descriptive statistics is data analysis, which is supported by more complex and comprehensive descriptive tools (Cobb & Moore, 1997). Recognizing the significance of data analysis, Libman (2010) proposed that utilizing real-life scenarios as the foundation for descriptive statistics assignments can facilitate students' comprehension of the purpose behind data analysis.

Inferential statistics extends beyond the data to make inferences about a broader population, acknowledging the omnipresence of variation and the inherent uncertainty in drawing conclusions (Moore, 2007). Kern (2014) highlighted the divergence in goals between descriptive statistics and

inferential statistics. Descriptive provides a quick overview to identify prominent patterns with minimal reliance on assumptions, while inferential statistics relies more heavily on the initial premises. Learning descriptive statistics is an essential first step in conducting research and should always come before making inferential statistical comparisons. This is because descriptive statistics enable the organized summarization of data by describing the relationship between variables in a sample or population (Makar & Rubin, 2018).

Researchers use inferential statistics to make predictions or inferences based on the data. By using inferential statistics, researchers can take data from samples and draw generalizations about a larger population (Baral, 2013). It is thus common for scholars to integrate both descriptive and inferential approaches in their study designs (e.g., McTavish & Loether, 2018).

Students' Difficulty and Misconceptions of Inferential Statistics

Statistical inference has proved to be challenging for most students (Makar & Rubin, 2018; Park, 2018). Students frequently struggle to comprehend the fundamental concepts or accurately interpret the outcomes of computation (Case & Jacobbe, 2018; Chance et al., 2004), for example, on sampling distribution.

Sampling distribution serves as the key to unlocking comprehension of statistical inference (Aguinis & Branstetter, 2007; Garfield et al., 2008; Setyani & Kristanto, 2020). However, misuse and misunderstanding of sampling distribution are frequently observed in the field of statistics (Lewis, 1999). Although students are typically introduced to probability distributions before sampling distributions, comprehending the idea of a distribution and making probabilistic interpretations have proved to be challenging during the learning process of inferential statistics (Kula & Koçer, 2020).

Students could have difficulties in comprehending the concepts of sample and distinguishing between sample statistics and population parameters (Garfield et al., 2008; Kula & Koçer, 2020). Their understanding of the variability of a sample statistic is usually restricted to different values obtained from drawing different samples, showing a lack of recognition of variability extending to the concept of distribution (Saldanha & Thompson, 2002). Equating a sampling distribution with a distribution of a sample, and confusing it with the distribution of raw data from the population is also common (Lipson, 2002; Yu & Behrens, 1994). Many students fail to comprehend the asymptotic behaviour of sampling distribution when repeated samples are drawn from a given population (Kula & Koçer, 2020).

The Central Limit Theorem (CLT), which forms the foundation for pursuing further comprehension and application of inferential statistics (Yu & Behrens, 1995), is also difficult for most students (Kim, 2020). Students often lack a thorough understanding of the sample size and distribution of the population which determine the normality of a sampling distribution. (Chance et al., 2004; Yu & Behrens, 1995; Zhang et al., 2022).

The rationale of hypothesis testing is not intuitive to many students and thus presents a major barrier to understanding statistical inference (Cobb & Moore, 1997; Liu & Thompson, 2009). The misconceptions that hypotheses can be "logically proved" (Vallecillos & Holmes, 1994), and the rejection of null hypothesis implies the truth of alternative hypothesis (Travers et al., 2017) are also common.

Significance level is often misinterpreted as the probability that the null hypothesis is true once the decision to reject it has been made (Batanero, 2000). Haller and Krauss (2002) showed two distinct types of misunderstandings regarding the significance level: (1) erroneous interpretation as random-percentage, such as for 5% significance level meaning that "it means, that the measure lies 5% above the random-percentage" (p. 2); (2) significance test enables an evaluation of the probabilities of hypotheses. Other difficulties related to significance level include the relation of sampling distribution to the level of significance, and the critical region (Vallecillos, 1999).

The p-value is commonly misinterpreted as the probability of the null hypothesis being true or false (McShane et al., 2019; Smith, 2018; Wright, 2002). Such a misinterpretation disregards the conditional probabilistic nature of the p-value (Gagnier & Morgenstern, 2017). Studies have also shown the misconception that the p-value represents the probability that the null hypothesis is true, given certain data (Badenes-Ribera et al. 2016; Krzywinski & Altman, 2013; McShane et al., 2019; Smith, 2018; Verdam et al. 2013), or the chance of getting the outcomes they observed from an experiment (Travers et al., 2017).

Students must comprehend and establish connections among numerous abstract concepts, such as the sampling distribution, significance level, null and alternative hypotheses, the test statistic and the p-value (Liu & Thompson, 2009; Sotos et al., 2007). These concepts are mutually connected, further making hypothesis testing difficult to comprehend (Emmert-Streib & Dehmer, 2019).

Objective and Research Question that Guides This Study

In this study, we aim to explore the understanding of inferential statistics among beginning undergraduate students in Singapore, in order to provide insight for educators to examine their instruction in pre-university statistics. Formal inferential statistics occupies a significant portion of the Singapore pre-university mathematics syllabus, even though informal inference (such as intuitive ideas of sampling) is not introduced at the secondary or primary levels. Notably, the CLT is included in Singapore's pre-university mathematics curriculum, while it is typically introduced at the university level in most other countries due to its abstract nature. These factors highlight the importance of assessing students' understanding of the concepts they learned in pre-university. We believe that the study we present here is timely as statistics is gaining wider recognition, and currently there is a scarcity of studies examining Singapore students' understanding of concepts of inferential statistics. Despite statistics receiving much attention in Singapore, studies in Singapore on statistics education are rare. A

search on Google Scholar and the ERIC database did not reveal any research conducted in Singapore on this specific topic. The research question (RQ) guiding this study is: "What are the common misconceptions held by beginning undergraduate students, who enrolled in mathematics courses in pre-university institutions in Singapore, regarding the concepts of sampling distribution and hypothesis testing in inferential statistics?"

Existing Tools to Measure Students' Understanding of Statistics

Researchers have developed various tools for assessing students' understanding of statistics. Cohen and Chechile (1997) introduced an interactive program to exclusively test conceptual understanding of probability distribution. Questions to evaluate students' comprehension via graphical representations were used. They emphasized the importance of open-ended questions in identifying unanticipated misconceptions, which provide clues to conceptual confusions.

Garfield (1998) created an instrument to assess students' statistical reasoning skills and understanding of probability and statistics concepts. Multiple-choice questions featuring statistics and probability problems were used. Students were required to select the response that best matches their own thinking and reasoning. The selection provides insights into students' thought processes and captures their reasoning behind their choices.

According to Turegun and Reeder (2011), multiple-choice items might restrict students' individual thinking processes and risk students reverse-engineering correct answers. Aiming for an in-depth examination of conceptual understanding, they used an open-ended questionnaire format which requires students to justify their answers.

The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project was developed to assist statistics teachers in creating and utilizing better assessments (Garfield & delMas, 2010). It offers an extensive online item database with over a thousand assessment items in three formats (open-ended, multiple-choice, performance tasks). These items provide resources to evaluate students' statistical literacy, reasoning, and thinking.

METHODS

This research is a case study of the misconceptions related to inferential statistics among mathematics pre-service teachers in Singapore. Even though this study involved only nine student teachers (this group of nine students was an entire cohort of the teacher education course from the degree program), we focused on deeper insights into the categories of students' misconceptions. Our study will serve as a foundation for more extensive future research, providing a focused perspective on the complexities of students' understanding of inferential statistics.

Case study enables researchers to explore a phenomenon through a diverse array of perspectives and offers the flexibility to deeply investigate intricate experiences and situations that may lack a definite or singular conclusion (Lucas et al., 2018). Through case study, the researcher offers a substantial depiction of the research context, thus enabling a comprehensive understanding of the core subject through detailed observation and examination (Creswell, 2015). Case study methodology is useful for investigating a contemporary phenomenon in which researchers have limited control over behavioural events (Yin, 2018).

Developing the Instrument

Examining the assessment tools described in the previous section, none of them fully met our research objective. Hence, we constructed our own instrument. We used a combination of both multiple-choice questions (three items) and open-ended questions (four items) in this study, modelled after the structures of the assessment tools described above.

We designed three multiple-choice questions to assess students' understanding of sampling distributions, the CLT, and the p-value. Four open-ended questions were designed to examine their understanding of the null and alternative hypothesis, significance level, test statistics, and critical regions. The selection of this mixed format stemmed from our consideration of the affordances of these two formats and the assessment tools we reviewed above.

For concepts such as sample distribution, CLT, and p-value, we identified various misconceptions reported in existing research studies from other educational systems and settings. Based on these findings, we aimed to examine whether Singaporean students in our study exhibited similar misconceptions. To assess this, we incorporated these misconceptions as distractors in multiple-choice questions. Multiple-choice questions are effective for assessing comprehension (Xu et al., 2016), allowing assessors to target specific aspects (Scharf & Baldwin, 2007). Our literature review confirmed that diverse misconceptions related to these concepts have been observed across different educational contexts (e.g., Haller & Krauss 2002; Kim, 2020; Kula & Koçer, 2020; Lewis, 1999; McShane et al., 2019; Vallecillos,1999).

Open-ended questions were used to obtain a more comprehensive assessment of students' comprehension of the null and alternative hypothesis, significance level, test statistics and critical regions. Since few misconceptions related to these concepts have been identified in other educational systems and settings, there was limited scope for creating effective multiple-choice questions. Thus, the use of open-ended questions can provide the advantage of uncovering unanticipated misconceptions (Turegun & Reeder, 2011).

Inferential reasoning is based on the idea that a sample, though only a part of the population, can still provide meaningful insights about the entire population. The process typically begins by forming a hypothesis about the situation, starting with a null hypothesis which is under the assumption that

chance is the only explanation (Batanero & Borovcnik, 2016). The null hypothesis states that the observed results are not significantly different from what would be expected by random variation (Batanero & Borovcnik, 2016; Moore et al., 2008). In contrast, the alternative hypothesis is a statement that contradicts the null hypothesis, suggesting that there is a real effect or difference (Moore et al., 2008). The observed data is just one of all possible samples that could be drawn from the population. The distribution of sample statistics (e.g. the sample mean in this study) across all possible samples of a specified size reflects the variability and is referred as the sampling distribution (Batanero & Borovcnik, 2016). According to the Central Limit Theorem, regardless of the population's original distribution, as long as the population has a finite standard deviation, the sampling distribution of the sample mean will approximate a Normal distribution when the sample size is sufficiently large (Moore et al., 2008; Sotos et al., 2007).

To draw conclusions, we assess how unusual the observed data is, or how extreme it is, assuming the null hypothesis is true. The p-value is the probability of obtaining the observed value (i.e. the test statistic in this study) or a more extreme value, under the null hypothesis (Batanero & Borovcnik, 2016; Moore et al., 2008). A smaller p-value indicates stronger evidence against the null hypothesis The significance level serves as a threshold to determine whether the data are consistent with the null hypothesis. If the p-value is equal to or smaller than the significance level, the data are deemed statistically significant, leading to a rejection of the null hypothesis (Batanero & Borovcnik, 2016; Moore et al., 2008). If a value of significance level is chosen, the critical value is a threshold or cutoff point that separates the region under the Normal distribution curve where the null hypothesis is rejected from the region where it is not rejected (Moore et al., 2008). The critical region is the range of values where the null hypothesis is rejected. The instrument and rationale of the items are given below:

Question 1

Multiple-choice question on sampling distribution: *Which of the following describes a sampling distribution of sample mean?* This question focuses exclusively on the sampling distribution of the sample mean because, in Singapore's pre-university mathematics curriculum, students are only introduced to the concept of sampling distribution through the sample mean. The same rationale applies to Question 2 below.

1. Option (a): The distribution of individual data points in a sample.
   This option inaccurately equates the sampling distribution with the distribution of raw data from the population. The objective of this option is to identify students' confusion between sampling distribution and the distribution of raw data (Lipson, 2002; Yu & Behrens, 1994).
2. Option (b): The distribution of sample mean calculated from all possible samples.
   This option is the correct answer (Batanero & Borovcnik, 2016; Garfield et al., 2008).

3. Option (c): The distribution of population mean.
   This option serves to identify students' confusion between the sample statistic and population parameter (Garfield et al., 2008; Kula & Koçer, 2020). Sample statistics are random variables with a probability distribution, while population parameters are constants.
4. Option (d): The distribution of a single sample.
   This option identifies students' inability to distinguish between sample distributions and sampling distributions (Lipson, 2002; Yu & Behrens, 1994). Sampling distributions, which indicate the distribution of a sample statistic (such as sample mean, sample proportion, etc.) calculated from all possible samples of identical size (Garfield et al., 2008; Lipson 2003), are distinct from sample distributions, which represent the distribution of a sample (Garfield et al., 2008; Lipson, 2002; Yu & Behrens, 1994).
5. Option (e): None of the above. (Please state your reasons below.)
   This option was introduced to assist us in identifying potential misconceptions which had not been anticipated by us. This option was introduced for the other multiple-choice questions with the same objective.

Question 2

Multiple-choice question on CLT: *The Central Limit Theorem states that:*

1. Option (a): The sample is approximately normally distributed if the sample size is large.
   This option inaccurately confuses the sampling distribution with the sample distribution (Lipson, 2002; Yu & Behrens, 1994), allowing us to identify the misconceptions discussed in option (d) of Q1 in the context of the CLT. In addition, this option enables us to assess whether students had an accurate understanding of the CLT, which exclusively guarantees the normality of sample statistics.
2. Option (b): The sample mean is approximately normally distributed in samples of any size.
   The CLT does not guarantee that the sample mean will be normally distributed in any sample size. This option aims to discern whether students lack the understanding of the sample size that determines the normality of a sampling distribution (Chance et al., 2004; Zhang et al., 2022).
3. Option (c): The sample mean is approximately normally distributed if the sample size is large enough.
   This option is the correct answer Sotos et al., 2007).
4. Option (d): The sample mean is approximately normally distributed in samples of any size if the population follows a normal distribution.
   This statement is true but not because of CLT, which does not require the population to follow a normal distribution for the sample mean to be approximately normally distributed. However,

students often lack a thorough understanding of how the distribution of the original population influences the normality of a sampling distribution (Chance et al., 2004; Yu & Behrens, 1995; Zhang et al., 2022). Thus, this option serves to examine this potential misconception.

5. Option (e): None of the above. (Please state your reasons below.)

Question 3

Multiple-choice question on p-value: *What is the p-value in hypothesis testing?*

1. Option (a): The p-value is the probability that the null hypothesis is true.
   This option identifies a prevalent misconception among students who erroneously associate the p-value with the probability that the null hypothesis is true (Badenes-Ribera et al., 2016; Krzywinski & Altman, 2013; McShane et al., 2019; Smith, 2018; Verdam et al., 2013).
2. Option (b): The p-value indicates the probability of obtaining the observed test statistic, assuming the null hypothesis is true.
   This option is the correct answer (Batanero & Borovcnik, 2016).
3. Option (c): The p-value quantifies the probability of getting the outcomes observed from an experiment.
   This option reflects a misconception among students who mistakenly perceive that the p-value is the probability of getting the outcomes observed from an experiment (Travers et al., 2017).
4. Option (d): The p-value represents the probability of mistakenly rejecting the null hypothesis when the null hypothesis is true.
   This option reflects a misconception among students who wrongly perceive the p-value as representing the probability of committing an error while rejecting the null hypothesis when the null hypothesis is true (i.e., type I error) (Sotos et al., 2009).
5. None of the above. (Please state your answer below.)

Question 4

Open-ended question on null and alternative hypothesis: *How do you formulate the null hypothesis and alternative hypothesis?* Many studies have shown that students frequently have difficulty transforming contextual information from word problems (e.g., Holling et al., 2008; Pape, 2004). This question is designed to assess students' understanding of the process of formulating hypotheses. By asking students to explain how they construct both the null and alternative hypotheses, we seek to understand how students formulate the hypotheses so that unveil possible misconceptions.

Question 5

Open-ended question on significance level: *Explain what you understand by significance level*. The significance level is typically provided in practice and examination questions on hypothesis testing. Students might use it to make decisions without fully grasping its concept. For instance, some students in the study conducted by Haller and Krauss (2002) interpreted the significance level as "the measure lies 5% above the random-percentage" (p. 2), exemplifying instances where certain students gave meaningless explanations of significance level. This question aims to gauge students' comprehension of the significance level. It prompts them to articulate the significance level, the response to which will elucidate the extent of students' comprehension regarding the meaning and function of the significance level.

Question 6

Open-ended question on test statistic: *How are test statistics used to assess the evidence against the null hypothesis? Explain.* Common classroom experience shows that students frequently resort to calculators for the computation of test statistics. The built-in statistical programs on graphing calculators automatically furnish both test statistics and p-values without the need for explicit calculation procedures. The passive engagement with dialogue boxes in statistical software programs may lead to weak comprehension of connections between test statistics and p-values. This question prompts students to explain how the test statistics are utilized to evaluate the evidence against the null hypothesis. By requiring a detailed explanation, it enables a comprehensive assessment of students' grasp of the significance and application of test statistics.

Question 7

Open-ended question on critical region, significance level and decision-making: *(a) How does the critical region relate to the significance level? (b) Hence, how is it used in the decision-making process in hypothesis testing?* In contrast to the automated reporting of test statistics and p-values by built-in statistical programs on graphing calculators, deriving corresponding critical values using graphing calculators requires employing the embedded statistical distribution, demanding a robust grasp of statistical and probabilistic principles. Thus, it is not surprising that passive learning was noted by anecdotal class evidence that students tend to avoid involvement with critical regions in hypothesis testing. This avoidance likely reduces their exposure to this concept and may result in a limited understanding of how this process relates to the concept of 'significance' and the underlying rationale guiding decision-making.

Part (a) of Question 7 aims to explore students' understanding of the interrelationship between critical regions and the significance level. Part (b) serves to explore their understanding of the role of critical regions in the decision-making process in hypothesis testing.

The Participants

The participants of this study were an entire cohort of nine mathematics pre-service teachers from the Singapore National Institute of Education from the first-year degree program in the university. They had completed their pre-university education, hence were familiar with the pre-university inferential statistics of the mathematics curriculum.

This cohort of pre-service teachers, also the first-year undergraduate mathematics students, had not yet pursued further statistics-related courses in the university. Their exposure to inferential statistics remained limited to their prior experiences in pre-university, rendering them suitable candidates for this study. Particularly, the majority were enrolled in different pre-university institutes in Singapore. This study was approved by the University's Ethics Review Committee. The invited pre-service teachers agreed to participate. They were provided with a hard copy of the questionnaire consisting of the seven questions described above. The survey, with 30 minutes duration, was conducted under usual closed-book conditions.

RESULTS AND DISCUSSION

Multiple-Choice Items

As shown from the students' responses to the multiple-choice items (Q1, Q2 and Q3), most students could answer the items on sampling distribution and the CLT correctly, while more than half the students showed a range of misconceptions of the p-value. This aligns with the findings from several other studies, which unveiled a greater diversity of misconceptions of the p-value, compared to misconceptions regarding the sampling distribution and the CLT (e.g., Gagnier & Morgenstern, 2017; McShane et al., 2019; Smith, 2018; Sotos et al., 2009; Wright, 2002). The students' responses to the three items are shown in Table 1. The correct answer for each question is highlighted in bold.

Table 1. The participants responses to the multiple-choice questions

Question	Option	The number of students who chose this option
Q1: Which of the following describes a sampling distribution of sample mean?	(a) The distribution of individual data points in a sample.	2
	(b) The distribution of sample mean calculated from all possible samples.	**6**
	(c) The distribution of population mean.	1
	(d) The distribution of a single sample.	0
	(e) None of the above. (Please state your reasons below.)	0
Q2: The Central Limit Theorem states that	(a) The sample is approximately normally distributed if the sample size is large.	2
	(b) The sample mean is approximately normally distributed in samples of any size.	0

Question	Option	The number of students who chose this option
	(c) The sample mean is approximately normally distributed if the sample size is large enough.	**7**
	(d) The sample mean is approximately normally distributed in samples of any size if the population follows a normal distribution.	0
	(e) None of the above. (Please state your reasons below.)	0
Q3: What is the *p*-value in hypothesis testing	(a) The p-value is the probability that the null hypothesis is true.	2
	(b) The p-value indicates the probability of obtaining the observed test statistic, assuming the null hypothesis is true.	**4**
	(c) The p-value quantifies the probability of getting the outcomes observed from an experiment.	0
	(d) The p-value represents the probability of mistakenly rejecting the null hypothesis when the null hypothesis is true.	3
	(e) None of the above. (Please state your answer below.)	0

In response to Question 1 (Q1), six students chose the correct option (b). Two students selected option (a), reflecting the misconception that sampling distribution refers to the distribution of raw data (Lipson, 2002; Yu & Behrens, 1994). One student chose option (c), showing the misunderstanding of a sampling distribution of sample mean as a sampling distribution of population mean (Garfield et al., 2008; Kula & Koçer, 2020). None of the students chose option (d), i.e., none equated the sampling distribution with the sample distribution (i.e., the distribution of a single sample) (Lipson, 2002; Yu & Behrens, 1994).

In Q2, seven students chose option (c), correctly identified that the CLT guarantees the sample mean with an approximate normal distribution provided the sample size is sufficiently large. Two students chose option (a), demonstrating a confusion of the sampling distribution with the sample distribution (Lipson, 2002; Yu & Behrens, 1994) within the context of the CLT. They recognized the necessity of a large sample size for the CLT to hold but might not have known the normality refers to the sample statistics rather than the sample itself. Upon a closer examination, student 2 was likely confused between sample statistics and population parameters in the response to Q1 by choosing option (c). Student 7 showed a misconception of the sampling distribution by perceiving it as the distribution of raw data in Q1 by incorrectly choosing option (a). These two students who consistently made the same mistakes in Q1 and Q2 showed their lack of understanding of sampling distribution and the CLT. The remaining students demonstrated a comparatively stronger comprehension of sampling distribution and the CLT.

None of the students chose option (b) in Q2, showing a lack of evidence of the misconception of the sample size requirement which is outlined by the CLT to guarantee the normality of the distribution of a sample mean (Chance et al., 2004; Zhang et al., 2022). No students selected option (d), showing no evidence of a confusion in understanding the influences of the distribution of the original population on the normality of a sampling distribution (Chance et al., 2004; Yu & Behrens, 1995; Zhang et al., 2022).

For Q3, four students chose the correct option (b). Two students (3 and 5) chose option (a), reflecting a misunderstanding of the p-value as the probability that the null hypothesis is true (Badenes-Ribera et al., 2016; Krzywinski & Altman, 2013; McShane et al., 2019; Smith, 2018; Verdam et al., 2013). Three students (1, 7 and 8) chose option (d), reflecting the misinterpretation of the p-value as probability of committing type I error (Sotos et al., 2009), that is, the probability of mistakenly rejecting the null hypothesis when the null hypothesis is true. None chose option (c), showing there was no evidence of the misconception that the p-values to the probability of getting the outcomes observed from an experiment (Travers et al., 2017) in this group of students.

Open-Ended Items

The students' responses to the four open-ended questions (Q4 to Q7) exhibited their weak understanding of the concepts related to hypothesis testing compared to sampling distribution and CLT from the multiple-choice questions. Examining the students' responses to Q4, we grouped students' understanding into four categories: (h1) The null hypothesis (or alternative hypothesis) is the statement that needs to be proved as true (or false, respectively). (h2) The alternative hypothesis (or null hypothesis) is the statement that needs to be proved as true (or false, respectively). (h3) The null hypothesis and alternative hypothesis represent two "opposite" statements. (h4) Decision-making with p-values and critical regions could be made without reference to the null or alternative hypotheses.

Categories (h1) and (h2) reveal the misconceptions that either the null or alternative hypothesis needs to be proved to be true. As described by Moore (1990): "Statistical significance is a way of answering the question 'Is the observed effect larger than can reasonably be attributed to chance alone?'" (p. 132). Hypothesis testing is about deciding which hypothesis is to be preferred from the information of the data, with the hope of controlling the probability of making type I error (Christensen, 2005; Hacking, 2016; Lenhard, 2006).

Category (h3) reflects the view that the alternative hypothesis is 'opposite' to the null hypothesis. Note that in one-tailed parametric hypothesis tests, the two hypotheses are not 'opposite' to each other ($H_0: \mu=\mu_0$ vs $H_1: \mu<\mu_0$ or $H_0: \mu=\mu_0$ vs $H_1: \mu>\mu_0$). Categories (h1), (h2) and (h3) share a common misconception that the purpose of hypothesis testing is either to prove the null (or alternative) hypothesis. Category (h4) is a procedural description of hypothesis testing by connecting the hypotheses, p-values and critical regions without providing conceptual understanding.

From the students' responses to Q5, their understanding of significance level was categorized into four categories: (s1) It is the probability that hypotheses (either null or alternative) are true or not true. (s2) It is employed in the decision of rejecting the null hypothesis. (s3) It is used to determine the critical region or critical value. (s4) It shows how significant the sample mean is to represent the population mean, only student 3 exhibited this comprehension.

Category (s1) demonstrated a misconception of the significance level being the probability of hypotheses being true or false. Category (s2) identified the significance level serving as a decision criterion. Students in category (s3) recognized that the role of significance level in finding the critical region or critical value is essential in making decisions. The student with the comprehension in category (s4) had likely not understood the role of significance level in hypothesis testing.

For Q6, only two students demonstrated an understanding of the role of test statistics in the decision-making of hypothesis testing. Their responses exhibited an understanding of the assessment aimed at verifying if observed value of test statistics lie within the critical region. Other students' responses exhibited misconception of test statistics. These misconceptions of test statistics were classified into three categories: (t1) incorrect responses due to the inability to recognize the connection between test statistics and p-value and critical region. It includes such responses as 'test statistics is calculated using probability and sample mean of the sample'; (t2) inability to identify the utilization of test statistics in the decision-making process with a critical region; and (t3) incorrect response alluding to test statistics serve to prove the truth of the null (or alternative) hypothesis. Category (t3) is similar to (h1) and (h2) of Q4, which is the misunderstanding that the purpose of hypothesis testing is to prove or disprove the null or alternative hypothesis. Students 7 and 9 in category (t3) did not explicitly connect the concept of test statistics with p-values, p-value was mentioned in their responses. Student 2 in category (t2), who mentioned the concept of critical region, did not indicate explicitly the relation between test statistics and these two concepts.

For Q7, only student 1 exhibited a grasp of both aspects regarding critical regions: the relationship between critical regions and significance level, i.e., critical regions are determined by significance level, as discussed in part (a), and the utilization of critical regions in the decision-making process, i.e., assessing if observed test statistics falls within critical regions, as assessed in part (b). The responses from students 2 to 9 exhibited misconceptions regarding the critical region were categorized into four categories: (c1) Critical regions are ranges in which the null hypothesis is true or false. (c2) Critical regions are intervals where most of the data will fall. (c3) The use of critical regions in the decision-making process involves assessing if p-values fall within critical regions. (c4) The utilization of critical regions in the decision-making process involves comparing the value of critical region with significance level. Student 6 held this misconception.

Similar to the misconception identified in Q4, where students mistakenly believed that the null hypothesis or the alternative hypothesis can be proved or disproved, category (c1) of Q7 reflects a misunderstanding of the critical regions as ranges for proving or disproving the null hypothesis.

Category (c2) reflects a misunderstanding between the distinction between the concept of data and random variables. Critical regions do not provide ranges for data but produce intervals for random variables. Students in categories (c3) and (c4) used p-values instead of critical regions.

The correctness of students' responses to the questions is summarized in Table 2. A checkmark symbol (✓) in the cell indicates that the student had demonstrated an understanding of the concepts assessed in the question. A cross-mark symbol (✗) in the cell represents that the student did not exhibit an understanding of the concepts assessed by the question.

Table 2. Tabulation of the correct responses of the participants

	Q1	Q2	Q3	Q4	Q5	Q6	Q7 (a)	Q7 (b)
The concept assessed	Sampling distribution	The central limit theorem	*p*-value	Null and alternative hypothesis	Significance level	Test statistics	How is the critical region related to the significance level	The critical region in decision-making process
S1	✗	✓	✗	✗	✓	✗	✓	✓
S2	✗	✗	✓	✗	✗	✗	✓	✗
S3	✓	✓	✗	✗	✗	✓	✗	✗
S4	✓	✓	✓	✗	✓	✓	✓	✗
S5	✓	✓	✗	✗	✗	✗	✗	✗
S6	✓	✓	✓	✓	✓	✗	✗	✗
S7	✗	✗	✗	✗	✓	✗	✓	✗
S8	✓	✓	✗	✗	✗	✗	✓	✗
S9	✓	✓	✓	✗	✓	✗	✗	✓

In considering the multiple-choice questions Q1 to Q3, only three students (4, 6, and 9) identified the correct description of sampling distribution, CLT and p-value. Another three students (3, 5, and 8) exhibited an understanding of sampling distribution and CLT in the multiple-choice questions. Two students (1 and 2) demonstrated understanding in only one question. One student (7) failed to correctly identify any of the concepts of the three questions. Most students gave the correct responses to the items on CLT (seven out of nine) and sampling distribution (six out of nine). Only four students gave the correct response to item 3 on p-value.

Upon closer scrutiny, the students' answers to the questions also exhibited misconceptions of concepts not directly addressed by the questions. We extracted the various misconceptions related to hypothesis testing and sampling distribution, and tabulated the number of students with these misconceptions (Table 3).

Table 3. Tabulation of the misconceptions of inferential statistics among the participants.

Misconception	Number of students
The sampling distribution is the distribution of raw data.	2
Recognizing the sampling distribution of the sample mean as the sampling distribution of the population mean.	1
The Central Limit Theorem guarantees the normality of sample distribution when the sample size is large enough.	2
The *p*-value is the probability that the null hypothesis is true.	2
The *p*-value is the probability of mistakenly rejecting the null hypothesis when the null hypothesis is true.	3
The null and (or) alternative hypothesis can be proved or disproved.	9
The significance level is the probability that hypotheses (either null or alternative) are true or not true.	3
Test statistics can be calculated by using probability.	1
Test statistics serve to prove the truth of the null or alternative hypothesis.	3
Critical regions are ranges in which the null hypothesis is true or false.	3
Critical regions are intervals where most of the data will fall.	1
The use of critical regions in the decision-making process involves assessing if *p*-values fall within critical regions.	3
The utilization of critical regions in the decision-making process involves comparing the value of critical region with significance level.	1

The most significant misconception among all the students was that the objective of hypothesis testing is to prove or disprove the null or alternative hypothesis.

Discussion

The students in this study exhibited misconceptions that are similar to those identified in other educational contexts related to sampling distribution, the CLT, and p-value (e.g., Garfield et al., 2008; Lipson, 2002; Sotos et al., 2009; Yu & Behrens, 1994; Zhang et al., 2022). They seem to have a better understanding of sampling distribution and CLT compared to concepts involved in hypothesis testing. We further revealed unanticipated misconceptions regarding the null and alternative hypothesis, significance level, test statistic, and critical region. These misconceptions offer insights into the challenges faced when learning about inferential statistics in pre-university mathematics in Singapore.

The misconception of the null or alternative hypothesis as being provable could be an indicator of a lack of statistical thinking due to their long-held grounding in mathematical thinking. Understanding the omnipresent nature of variation within data is a core element of statistical thinking (Engel & Sedlmeier, 2005; Moore, 1990; Snee, 1990; Wild & Pfannkuch, 1999). Students who mistakenly claim that a hypothesis can be proved as true or false exhibited a deficiency in understanding the variation in empirical data.

Deterministic thinking, which mathematics is usually associated with, does not attempt to align the variability in data with a probability model (Pfannkuch & Wild, 2004). Determinism adopts a view of truth and knowledge that is absolute, suggesting that an outcome can be predicted or determined (Serradó et al., 2005). Students are often exposed to questions that require them to prove or disprove mathematical statements or expressions in learning mathematics. Mathematical computations commonly involve deterministic algorithms. Hence, the result is invariably either true or false. In mathematics education, students are frequently guided towards making deterministic generalizations that apply to a variety of problems (Groth, 2013), quite contrary to statistical thinking.

In many countries, elements of inferential statistics have been introduced into the secondary school mathematics syllabus, e.g., the International General Certificate of Secondary Education (IGCSE) mathematics syllabuses (Toh, 2023a; 2023b; 2023c). Pfannkuch and Wild (2015) recommended introducing inferential statistics gradually throughout the curriculum, rather than presenting them as a complex network of integrated ideas in the final years of schooling. Some form of inferential statistics, such as the notion of sampling, can be taught to students as early as the lower secondary levels (e.g., Toh et al., 2018; 2021).

CONCLUSION

This is the first study on the understanding of inferential statistics among mathematics pre-service teachers in Singapore. The findings of this study could serve as a foundation for further study on students' and teachers' knowledge of statistics. In particular, we have demonstrated that Singapore pre-service teachers share many misconceptions in inferential statistics as reported in many studies abroad.

Statistics is commonly integrated into the mathematics curriculum before tertiary studies in most of the education systems worldwide, including Singapore. The Singapore mathematics curriculum document does not make explicit reference to statistical thinking, statistical literacy or statistical reasoning, or the need for data or dealing with variation. It appears that the syllabus explicitly describes the deterministic computations associated with statistics. The lack of description of statistical thinking in the syllabus might likely have led teachers to overlook the significance of statistical thinking or conceptual understanding in statistics. This study also shows that the teachers had the tendency to describe statistical concepts using procedural steps without much conceptual understanding. Given the intricate and strongly interconnected nature of the fundamental concepts in inferential statistics, learners with weak understanding of concepts may find it challenging to grasp the logic of inferential statistics. We recommend that teachers dedicate sufficient attention to developing their learners' conceptual understanding when teaching inferential statistics. Conceptual understanding of statistics can best be developed through discussing the "commonsense" of the statistical procedures through engaging the learners in the real-world context of carrying out the statistical procedures.

This study also showed that the pre-service teachers overlooked the assumptions in performing hypothesis testing. We argue for the importance of checking the assumptions before hypothesis testing during instruction, as it serves to ensure the validity of performing hypothesis testing. This enables learners to see the process of hypothesis testing as a whole, with interconnecting key components such as null hypothesis, alternative hypothesis, test statistics, and p-value.

Students may better understand formal methods of statistical inference if they develop informal ideas of inference early in their studies. Delayed exposure to inferential statistics might be a potential factor contributing to students' struggle in grasping inference-related concepts. The exploratory study presented here aims to support a more inclusive society developing statistical literacy, hence integrating learners into the world involving data through improving statistics instruction.

Limitation of the Study

The limited availability of literature on misconceptions related to hypothesis testing led to the inclusion of different assessment formats (multiple-choice and open-ended questions) for different concepts, introducing a potential limitation to this study. While open-ended items allow participants to express their understanding, the written format may pose challenges on the level of detail provided.

Despite sharing a common pre-university curriculum, teacher quality, teaching approaches and institutional resources vary among different institution. These differences may impact students' understanding of inferential statistics, as well as their learning approaches. Taking an alternative perspective, these students were spread from different pre-university institutions, suggesting a 'balance' out the variation across these institutions. The students in this study represent the top performers in their pre-university cohort. Thus, their misconceptions are likely to be also prevalent among the general student population.

ACKNOWLEDGMENTS

The authors would like to thank the National Institute of Education and the Nanyang Technological University for the support during the period of postgraduate study by the first author.

DECLARATIONS

Author Contribution : QP : Conceptualization, data analysis, investigating and methodology, writing-original draft, wiring-review and editing.
TLT : Supervision, writing-review and editing.
YZ : Supervision, writing-review and editing.

Funding Statement : This study is not funded.

Conflict of Interest : The authors declare no conflict of interest

Additional Information : Additional information is available for this paper.

REFERENCES

Aguinis, H., & Branstetter, S. A. (2007). Teaching the concept of the sampling distribution of the mean. *Journal of Management Education, 31*(4), 467–483. https://doi.org/10.1177/1052562906290211

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01247.

Baral, K. K. (2013). Descriptive and inferential statistics. Retrieved from https://wbsche.wb.gov.in/assets/pdf/Political-Science/Descriptive-and-Inferential-statistics.pdf.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*(1–2), 75–97. https://doi.org/10.1207/s15327833mtl0202_4.

Batanero, C., & Borovcnik, M. (2016). *Statistics and probability in high school*. Sense Publishers. https://doi.org/10.1007/978-94-6300-624-8.

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education.* American Statistical Association.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3–15). Springer. https://doi.org/10.1007/1-4020-2278-6_1.

Braithwaite, D. W., & Sprague, L. (2021). Conceptual knowledge, procedural knowledge, and metacognition in routine and nonroutine problem solving. *Cognitive Science*, *45*(10). https://doi.org/10.1111/cogs.13048.

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in learning inference from a simulation-based approach. *Statistics Education Research Journal, 17*(2), 9–29. https://doi.org/10.52041/serj.v17i2.156.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). https://doi.org/10.1080/10691898.2002.11910677.

Chance, B., del Mas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). https://doi.org/10.1007/1-4020-2278-6_13.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*(2), 121–126. https://doi.org/10.1198/000313005x20871.

Cobb, G. P., & Moore, D. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly, 104*(9), 801–823. https://doi.org/10.1080/00029890.1997.11990723.

Cohen, S., & Chechile, R. A. (1997). Probability distributions, assessment and instructional software: Lessons learned from an evaluation of curricular software. In I. Gal & J.B. Garfield (Eds), *The*

assessment challenge in statistics education (pp. 253-262). https://iase-web.org/documents/book1/chapter19.pdf?1402524893.

Creswell, J. W. (2015). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research (fifth edition)*. Pearson.

Da Silva, A. A., Barbosa, M. A., De Souza Velasque, L., Da Silveira Barroso Alves, D., & Magalhães, M. N. (2021). The COVID-19 epidemic in Brazil: How statistics education may contribute to unravel the reality behind the charts. *Educational Studies in Mathematics, 108*(1–2), 269–289. https://doi.org/10.1007/s10649-021-10112-6.

Dani, A., & Quraan, E. A. (2023). Investigating research students' perceptions about statistics and its impact on their choice of research approach. *Heliyon*, *9*(10), e20423. https://doi.org/10.1016/j.heliyon.2023.e20423.

Delcham, H., & Sezer, R. (2010). Write-skewed: Writing in an introductory statistics course (EJ917149). ERIC. https://eric.ed.gov/?id=EJ917149.

Emmert-Streib, F., & Dehmer, M. (2019). Understanding statistical hypothesis testing: The logic of statistical inference. *Machine Learning and Knowledge Extraction, 1*(3), 945–961. https://doi.org/10.3390/make1030054.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1). https://doi.org/10.52041/serj.v16i1.213.

Engel, J., & Sedlmeier, P. (2005). On middle-school students' comprehension of randomness and chance variability in data. *ZDM – Mathematics Education, 37*(3), 168–177. https://doi.org/10.1007/s11858-005-0006-4.

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*. American Statistical Association.

Gagnier, J., & Morgenstern, H. (2017). Misconceptions, misuses, and misinterpretations of p values and significance testing. *Journal of Bone and Joint Surgery, American Volume, 99*(18), 1598–1603. https://doi.org/10.2106/jbjs.16.01314.

Garfield, J. B. (1998). The statistical reasoning assessment: Development and validation of a research tool. *Proceedings of the Fifth International Conference on Teaching Statistics* (pp.781-786). https://iase-web.org/documents/papers/icots5/Topic6u.pdf?1402524957.

Garfield, J. B., Delmas, R., Chance, B. (2003, April). The web-based ARTIST: Assessment resource tools for improving statistical thinking. Assessment of Statistical Reasoning to Enhance Educational Quality [Symposium]. AERA Annual Meeting, Chicago. https://www.causeweb.org/cause/archive/artist/articles/AERA_2003.pdf.

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer.

Garfield, J. B., & delMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics, 32*(1), 2–7. https://doi.org/10.1111/j.1467-9639.2009.00373.x.

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. https://doi.org/10.1080/10986065.2013.770718

Groth, R. E. (2015). Working at the boundaries of mathematics education and statistics education communities of practice. *Journal for Research in Mathematics Education, 46*(1), 4–16. https://doi.org/10.5951/jresematheduc.46.1.0004.

Hacking, I. (2016). *Logic of statistical inference*. Cambridge University Press.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1–20. https://bit.ly/49NsvJv.

Hannigan, A., Gill, O., & Leavy, A. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education, 16(*6), 427–449. https://doi.org/10.1007/s10857-013-9246-3.

Hijazi, R., & Shaqlaih, A. S. (2023). Statistical thinking at early primary school levels: Curriculum perspectives in UAE textbooks. *Statistics Education Research Journal, 22*(2), 13. https://doi.org/10.52041/serj.v22i2.447.

Holling, H., Blank, H., Kuchenbäcker, K., & Kuhn, J.-T. (2008). Rule-based item design of statistical word problems: A review and first implementation. *Psychology Science, 50*(3), 363–378.

Holmes, P. (2003). 50 years of statistics teaching in English schools: Some milestones. *The Statistician, 52*(4), 439–463. https://doi.org/10.1046/j.1467-9884.2003.372_1.x.

Kern, S. E. (2014). Inferential statistics, power estimates, and study design formalities continue to suppress biomedical innovation. ArXiv:1411.0919 [Stat]. https://arxiv.org/abs/1411.0919.

Khalid, M., & Embong, Z. (2019). Sources and possible causes of errors and misconceptions in operations of integers. *International Electronic Journal of Mathematics Education, 15*(2). https://doi.org/10.29333/iejme/6265.

Kim, S. (2020). A case study of a lesson on the sample mean for prospective mathematics teachers. *Mathematics, 8*(10), 1706. https://doi.org/10.3390/math8101706.

Krzywinski, M., & Altman, N. (2013). Significance, p values and t-tests. Nature Methods, 10(11), 1041–1042. https://doi.org/10.1038/nmeth.2698.

Kula, F., & Koçer, R. G. (2020). Why is it difficult to understand statistical inference? Reflections on the opposing directions of construction and application of inference framework. *Teaching Mathematics and Its Applications, 39*(4), 248–265. https://doi.org/10.1093/teamat/hrz014.

Lenhard, J. (2006). Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson. *The British Journal for the Philosophy of Science, 57*(1), 69–91. https://doi.org/10.1093/bjps/axi152.

Lewis, C. P. (1999). *Understanding the sampling distribution and the Central Limit Theorem.* (ED426100). ERIC. https://files.eric.ed.gov/fulltext/ED426100.pdf.

Libman, Z. (2010). Integrating real-life data analysis in teaching descriptive statistics: A constructivist approach. *Journal of Statistics Education, 18*(1). https://doi.org/10.1080/10691898.2010.11889477.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. *Proceedings of the Sixth International Conference on Teaching Statistics, 1-6*. https://iase-web.org/documents/papers/icots6/6c1_lips.pdf.

Lipson, K. (2003). The role of the sampling distribution in understanding statistical inference. *Mathematics Education Research Journal, 15*(3), 270–287. https://doi.org/10.1007/bf03217383.

Liu, Y., & Thompson, P. W. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies: An International Journal, 4*(2), 126–138. https://doi.org/10.1080/15544800902741564.

Lucas, P., Fleming, J., & Bhosale, J. (2018). The utility of case study as a methodology for Work-Integrated Learning research. *International Journal of Work-Integrated Learning, 19*(3), 215–222. http://files.eric.ed.gov/fulltext/EJ1196748.pdf.

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In: Ben-Zvi, D., Makar, K., Garfield, J. (Eds.), *International handbook of research in statistics education* (pp. 261-294). Springer. https://doi.org/10.1007/978-3-319-66195-7_8.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician, 73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253.

McTavish, D. J., & Loether, H. J. (2018). *Descriptive and inferential statistics*. SAGE Publications. https://doi.org/10.4135/9781071802656.n12.

Ministry of Education. (2007). *Introduction of Calculators in primary 5-6 mathematics*. Author.

Ministry of Education. (2013). *Primary One to Six Mathematics syllabus*. Author.

Ministry of Education. (2020a). *Pre-University Higher 2 Mathematics syllabus*. Author.

Ministry of Education. (2020b). *Secondary One to Four Express Course Mathematics syllabus*. Author.

Mishra, P., Pandey, C. M., Singh, U., Keshri, A., & Sabaretnam, M. (2019). Selection of appropriate statistical methods for data analysis. *Annals of Cardiac Anaesthesia, 22*(3), 297–301. https://doi.org/10.4103/aca.ACA_248_18.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). National Academies Press.

Moore, D. S. (2007). *The Basic practice of statistics (4th ed.).* W.H. Freeman.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2008). *Introduction to the practice of statistics (6th ed.).* W H Freeman.

Naseer, M. S. (2015). Analysis of students' errors and misconceptions in pre-university mathematics courses. *Proceedings of the 1st International Conference on Teaching Statistics* (pp. 34-39).

Page, W., & Moore, D. S. (1988). Should mathematicians teach statistics? *College Mathematics Journal, 19*(1), 2–7. https://doi.org/10.1080/07468342.1988.11973073.

Pape, S. J. (2004). Middle school children's problem-solving behavior: A cognitive analysis from a reading comprehension perspective. *Journal for Research in Mathematics Education, 35*(3), 187–219. https://doi.org/10.2307/30034912.

Park, R. (2018). Practical teaching strategies for hypothesis testing. *The American Statistician, 73*(3), 282–287. https://doi.org/10.1080/00031305.2018.1424034.

Pfannkuch, M., & Ben-Zvi, D. (2011). Developing teachers' statistical thinking. 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. 323–333). https://doi.org/10.1007/978-94-007-1131-0_31.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17–46). https://doi.org/10.1007/1-4020-2278-6_2.

Pfannkuch, M., & Wild, C.J. (2015). Laying foundations for statistical inference. In: S. J. Cho (Ed), *Selected regular lectures from the 12th international congress on mathematical education* (pp. 653–666). Springer. https://doi.org/10.1007/978-3-319-17187-6_36.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3). https://doi.org/10.1080/10691898.2002.11910678.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257–270. https://doi.org/10.1023/A:1023692604014.

Scharf, E. M., & Baldwin, L. P. (2007). Assessing multiple choice question (MCQ) tests - a mathematical perspective. *Active Learning in Higher Education, 8*(1), 31–47. https://doi.org/10.1177/1469787407074009.

Scheaffer, R. L. (2006). Statistics and mathematics: On making a happy marriage. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance* (pp. 309-321). National Council of Teachers of Mathematics.

Serradó, A., Azcárate, P., & Cardeñoso, J. M. (2005). Randomness in textbooks: the influence of deterministic thinking. *Proceedings of CERME 4: Fourth Conference of the European Society for Research in Mathematics Education* (pp. 1-10).

Setyani, G. D., & Kristanto, Y. D. (2020). *A case study of Promoting informal inferential reasoning in learning sampling distribution for high school students*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2002.04384.

Shi, N.-Z., He, X., & Tao, J. (2009). Understanding statistics and statistics education: A Chinese perspective. *Journal of Statistics Education, 17*(3). https://doi.org/10.1080/10691898.2009.11889538.

Smith, R. J. (2018). The continuing misuse of null hypothesis significance testing in biological anthropology. *American Journal of Physical Anthropology, 166*(1), 236–245. https://doi.org/10.1002/ajpa.23399.

Snee, R. D. (1990). Statistical thinking and its contribution to total quality. *The American Statistician, 44*(2), 116–121. https://doi.org/10.2307/2684144.

Sotos, A. E. C., Vanhoof, S., Van Den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98–113. https://doi.org/10.1016/j.edurev.2007.04.001.

Sotos, A. E. C., Vanhoof, S., Van Den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education, 17*(2). https://doi.org/10.1080/10691898.2009.11889514.

Stapor, K. (2020). Descriptive and inferential statistics. *Introduction to Probabilistic and Statistical Methods with Examples in R* (pp. 63–131). https://doi.org/10.1007/978-3-030-45799-0_2.

Toh, T. L. (2023a). *Cambridge lower secondary mathematics: Stage 7*. Marshall Cavendish Education.

Toh, T. L. (2023b). *Cambridge lower secondary mathematics: Stage 8.* Marshall Cavendish Education.

Toh, T. L. (2023c). *Cambridge lower secondary mathematics: Stage 9*. Marshall Cavendish Education.

Toh, T. L., Chan, C. M. E., Cheng, L. P., Lim, K. M., & Lim, L. H. (2018). Use of comics and its adaptation in the mathematics classroom. In *MATHEMATICS INSTRUCTION: GOALS, TASKS AND ACTIVITIES: Yearbook 2018, Association of Mathematics Educators* (pp. 67-85). Association of Mathematics Educators.

Toh, T. L., Cheng, L. P., Lim, L. H., & Lim, K. M. (2021). Teaching lower secondary statistics through the use of comics. In *MATHEMATICS—CONNECTION AND BEYOND: Yearbook 2020 Association of Mathematics Educators* (pp. 33-53). Association of Mathematics Educators.

Travers, J. C., Cook, B. G., & Cook, L. (2017). Null hypothesis significance testing and p values. Learning Disabilities Research and Practice, 32(4), 208-215. https://doi.org/10.1111/ldrp.12147.

Turegun, M., & Reeder, S. (2011). Community college students' conceptual understanding of statistical measures of spread. *Community College Journal of Research and Practice, 35*(5), 410–426. https://doi.org/10.1080/10668920903381854.

Vallecillos, J. A., & Holmes, P. (1994). Students' understanding of the logic of hypothesis testing. *Proceedings of the Fourth International Conference on Teaching Statistics.*

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. Bulletin of the International Statistical Institute: *Proceedings of the Fifty-Second Session of the International Statistical Institute, 58*, 201-204. https://www.stat.auckland.ac.nz/~iase/publications/5/vall0682.pdf.

Verdam, M. G. E., Oort, F. J., & Sprangers, M. a. G. (2013). Significance, truth and proof of p values: Reminders about common misconceptions regarding null hypothesis significance testing. *Quality of Life Research, 23*(1), 5–7. https://doi.org/10.1007/s11136-013-0437-2.

Vere-Jones. (1995). The coming of age of statistical education. *International Statistical Review, 63*(1), 3–23. https://doi.org/10.2307/1403774.

Weiland, T., Mojica, G., Engledowl, C., & Jones, R. S. (2019). *Statistics education: (Re)Framing past work for taking a holistic approach in the future* (ED606916). ERIC. https://files.eric.ed.gov/fulltext/ED606916.pdf.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x.

Wright, D. B. (2002). *First steps in statistics*. Sage Publications (CA).

Xu, X., Kauer, S. D., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology, 2*(2), 147–158. https://doi.org/10.1037/stl0000062.

Yin, R. K. (2018). *Case study research and applications: Design and methods (6th ed.).* Sage Publications.

Yu, C. H., & Behrens, J. T. (1994). Identification of misconceptions in learning statistical power with dynamic graphics as a remedial tool. *American Statistical Association Proceedings of the Section on Statistical Education,* 242-246. https://www.creative-wisdom.com/pub/power/power.pdf .

Yu, C. H., & Behrens, J. T. (1995). Identification of misconceptions in the Central Limit Theorem and related concepts and evaluation of computer media as a remedial tool (ED395989). ERIC. https://files.eric.ed.gov/fulltext/ED395989.pdf.

Zhang, X., Astivia, O. L. O., Kroc, E., & Zumbo, B. D. (2022). *How to think clearly about the Central Limit Theorem*. Psychological Methods. https://doi.org/10.1037/met0000448.

Zhang, Q., & Stephens, M. (2016). Teacher capacity as a key element of national curriculum reform in statistical thinking: a comparative study between Australia and China. In D. Ben-Zvi & K. Makar (Eds.), *The Teaching and Learning of Statistics* (pp. 301–313). https://doi.org/10.1007/978-3-319-23470-0_36.