Cognitive diagnostic assessment in university statistics education: Valid and reliable skill measurement for actionable feedback using learning dashboards

SUPPLEMENTARY MATERIAL

Contents

I Attribute descriptions

This section provides a detailed description of each of the 9 identified attributes with references to empirical studies that show the relevance of relevance of addressed concepts and the occurrence of misconceptions.

A. *Understanding center & spread.* Variability has a central role in inferential statistics. Unfortunately, several misconceptions in inferential statistics that are described by Castro Sotos, Vanhoof, Van den Noortgate, and Onghena (2007) stem from a lack of insight in the idea of variability in random events, which should be developed before the understanding of more complex concepts (delMas & Liu, 2005). Students often lack in-depth understanding of these earlier concepts, and many students complete their introductory statistics course without being able to integrate and apply these ideas (Chance, delMas, & Garfield, 2004). Therefore, this attribute is defined to evaluate students' insight in the variability of random events and understanding of varying observations. Students should be able to interpret and reason about common measures of center and spread (e.g., mean, median, range, standard deviation) and how data transformations and handling of outliers affect these measures.

B. *Interpreting univariate graphical representations.* This attribute refers to distributions of sample data that students learn to graph, describe, and interpret. Students should be able to describe distributions of data (shape, center, variability, and unusual observations) based on graphical representations. This requires familiarity with common univariate graphical displays, such as histograms, boxplots, dotplots, and bar charts. Literature shows that students experience difficulty in connecting measures of spread and graphical representations (Bakker & Gravemeijer, 2004; delMas, Garfield, Ooms, & Chance, 2007).

C. *Graphically comparing groups.* In statistics education there is much attention for the process of comparing two groups (Makar & Confrey, 2002). This requires students to consider not only measures of center, but also the issue of variability to evaluate whether differences in center between the groups are meaningful, i.e., to compare both variability between and within groups. The topic comparing groups can be structured as an informal version of statistical inference (Garfield & Ben-Zvi, 2008, Chapter 11), for example by learning students to make informal inferences about group differences based on boxplots. Misunderstandings regarding this topic may stem from students' difficulty to view data as an aggregate; as an entity with characteristics such as center, spread and shape (Konold, Pollatsek, Well, & Gagnon, 1997). Further, a persistant misconception in this regard is the belief that groups must be of equal size to make valid comparisons (delMas et al., 2007).

D. *Understanding sampling variability.* The key idea in statistical inference is that a sample provides incomplete information about the population from which it is drawn. A common misunderstanding relates to the law of large numbers, which states that as sample size increases, the sample mean will be closer to the population mean. Students seem to believe that both small and large samples are equally representative of the population they come

from, in other words, they believe in the law of small numbers (Tversky & Kahneman, 1971). This bias regarding insensitivity to sample size is part of the representativeness heuristic.

A fundamental concept in inferential statistics is sampling variability: not all samples are identical, so not all of them will resemble the population and sample statistics (such as the sample mean) will vary across samples. Sampling variability is central to statistical inference, yet conceptually difficult for students to grasp. This stems from the different levels of concreteness of the same concept in descriptive and inferential statistics (Schuyten, 1991). The unit of analysis in descriptive statistics is a single case, whereas in inferential statistics the unit of analysis is the sample. The sample is now considered as a unit from another population, namely the set of all possible samples with the given size. Further, students need to know that although sample statistics vary from sample to sample, they follow a predictable pattern where some values are more or less likely than others to be drawn from a population (delMas et al., 2007).

E. *Understanding sampling distributions.* Another key idea is that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of the distribution of that variable in the population (a direct result of the central limit theorem). Students need to distinguish between the (unknown) population mean, the sample mean, the possible values of different means that would be obtained with different samples of the given size, and the theoretical mean of these possible values, which corresponds to the population mean in case of random sampling (Batanero, Godino, Vallecillos, Green, & Holmes, 1994). Students may confuse the population and sampling distributions (Chance et al., 2004).

Batanero, Tauber, and Sánchez (2004) studied students' reasoning about the normal distribution, which is a complex idea containing many different elements to be taken into account when assessing students' understanding (e.g., definitions and properties, data representations, problem solving). They found that students experience difficulty differentiating between empirical almost-normal distributions and theoretical normal distributions. Whereas empirical distributions are used to describe and interpret the data and to think about models to explain variability in the data, theoretical distributions are models to fit the data, to explain variability, or to make predictions (Garfield et al., 2008). An occurring misconception related to this issue concerns the distinction between the real sampling distribution and a theoretical normal that is used as an approximation of this distribution used for null hypothesis significance testing. Students experience difficulties with this distinction (Castro Sotos et al., 2007). Another misconception is that students believe that the larger the sample size, the closer the distribution of the data resembles a normal distribution, even if the underlying population distribution is nonnormal (Bower, 2003). That is, students confuse the sample and sampling distributions (Lipson, 2002). This results in the belief that the theorem and the normal distribution can always be applied, leading to students' inability to justify its use.

F. *Understanding the standard error.* The standard error quantifies how much variability is expected across sample means and this information is used to make statistical inferences. Students should be able to explain the relation between the standard error, the population variance, and sample size. This involves understanding that statistics from small samples vary more than statistics from large samples. Students tend to neglect the effect of sample size on this variability (Chance et al., 2004). In addition, students experience difficulty in understanding how sampling error is used to make informal inference about a sample mean (delMas et al., 2007).

G. *Understanding principles of hypothesis testing.* Null hypothesis significance testing (NHST) is widely used for statistical inference, which is a blend of Neyman-Pearson's and Fischer's approaches to hypothesis testing and combines significance testing with decision making (Perezgonzalez, 2015). However, students sometimes tend to neglect the parallelism between hypothesis testing and decision processes (Castro Sotos et al., 2007). It is important that students learn to grasp the logic of significance tests and understand the the theoretical idea of finding evidence against a null hypothesis. Difficulties around significance levels and p-values (see attribute H) can directly result in misconceptions about the information that hypothesis tests provide and the conclusions that can be drawn from the tests. One misconception is that hypothesis tests are seen as mathematical proofs, i.e., that the results are deterministic and prove the null hypothesis to be true of false (Vallecillos, 2000). Another common fallacy is the belief that obtaining data that is unlikely given the null hypothesis implies that the null hypothesis is improbable, which is referred to as the *illusion of probabilistic proof by contradiction* (Falk & Greenbaum, 1995).

H. *Evaluating NHST results.* In NHST, after specifying hypotheses and performing significance tests, the obtained results are interpreted in terms of significance. This requires knowledge and understanding about significance levels and p-values. Many misconceptions in this process occur that relate to the distinction between the significance level and p-values and to their conditional nature (Haller & Krauss, 2002). The most common misconception is switching the terms in the conditional probabilities (Falk, 1986). In that case, the significance level is incorrectly interpreted as the probability that the null hypothesis is true once it is rejected, and the p-value is incorrectly interpreted as the probability that the null hypothesis is true given the observed data. This has been shown to be a persistent and deep misconception; see for example Vallecillos and Batanero (1997), Williams (1998), and Haller and Krauss (2002). Another type of misconception results from ignoring the conditional nature of significance levels and p-values, and hence considering those as single event probabilities. This can result in incorrectly interpreting the significance level as the probability of one hypothesis, or as the probability of making a mistake, or incorrectly interpreting the p-value as the probability that the event happened by chance (Castro Sotos et al., 2007). Haller and Krauss (2002) empirically show the occurrence of this type of misconception.

Beyond the misconceptions regarding the definition of p-values, the practical implications of results are often misunderstood. This involves incorrectly interpreting a statistically significant result as practically important (Gliner, Leech, & Morgan, 2002). It should be stressed that p-values do not provide measures of importance (Gagnier & Morgenstern, 2017). Statistical significance does not imply practical relevance, and neither the other way around; interpreting the practical relevance of results requires more detailed information about the research design and effect sizes. Students must acknowledge this and understand how different factors affect NHST results. This includes understanding the influences of effect size, population variance, sample size, and significance level on power. Based on this, students should be able to evaluate the quality and justification of evidence against the null hypothesis.

I. *Understanding and using confidence intervals.* Confidence intervals (CIs) are often argued to be a more useful alternative or complement to null hypothesis significance testing (e.g., Cumming & Finch, 2005; Fidler & Loftus, 2009; Reichardt & Gollob, 2016). However, the interpretation of CIs is not straightforward either. The key point is that a CI indicates a property of the performed procedure if used repeatedly, and not a property of the parameter obtained with the sample at hand (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Students tend to assign probabilities to parameters or hypotheses, which cannot be done within the frequentist framework. A common misconception is that CIs are viewed as plausible values of the sample mean rather than the population mean (Cumming, Williams, & Fidler, 2004), or, less frequently, as a range of individual scores (Fidler, 2006). These misunderstandings of what CIs represent can result in wrong interpretation of the comparison of CIs to compare independent means. Belia, Fidler, Williams, and Cumming (2005) found that participants widely believed that two 95% CIs with error bars that 'just touch' imply that means are just significantly different ($p < .05$). However, overlap in the CIs does not necessarily imply that there is no significant difference (Cumming & Finch, 2005). In addition to the interpretation of CIs, students learn how different factors influence the width of the CI (i.e., population variance, sample size, and confidence level). However, the effects of sample size and confidence level on the width of CIs are often misunderstood (Kalinowski et al., 2010).

II Interview procedures

This section provides detailed information about the participants, interview structure, and coding procedures of the think-aloud study.

1 Participants

- Participants were recruited via announcements during their second statistics course of their bachelor program. This means all participating students had successfully completed the introductory statistics course earlier.
- Students were all native Dutch. However, in their introductory statistics courses the reading materials (books, papers) were in English and students are expected to be proficient in English language.
- Participants received a 15 euro gift card as compensation.
- An information letter was sent to the students by e-mail and all students provided informed consent before participation.
- Only a subset of the 59 items was presented to each student. On average, students solved 27.5 items ($SD = 7.9$, $min = 16$, $max = 41$). We ensured each item was answered by at least 3 students by presenting different subsets in different orders.

2 Structure of the interview

The interviews took place via Microsoft Teams and were divided in three parts: introduction, main part, and reflection. This subsection describes the interview protocol for each part.

Introduction: 10 minutes

- Greet the student.
- Ask about the student's background in statistics:
  - Which educational program are they enrolled in?
  - Which statistics courses have they followed and successfully finished?
  - How would they rate their own performance in these courses: basic, good, very good or excellent?
- Explain the interview procedures to the student:
  - Items are presented one by one on a PowerPoint slide via screen sharing and the student is asked to start by reading it out loud.
  - The student narrates his/her entire thought process up to selecting the answer.
  - The statistics items are formulated in English, but students may think aloud in Dutch (their native language). If English terms are unknown to the student, the interviewer will translate these.

  - Other than that, the interviewer stays silent and does not provide feedback or clarifications; they only interrupts if the student stops speaking to remind the student to keep thinking aloud.
  - After approximately 60 minutes we proceed with reflection on the presented items. It does not matter how many items the student solves; there are no consequences. The student is encouraged to take all the time they need to answer an item and to explain reasoning.
  - The interviewer will make notes if the student's reasoning at a certain item was unclear. At the end of the interview clarifying questions can be asked to successfully ascertain the fine-grained cognitive processes that the students demonstrated as they solved the items.
  - At the end of the interview there is also time for the student to ask questions, for example about which answers are correct at specific items.
- Provide practical information:
  - The interview will be recorded (audio). In Microsoft Teams it is only possible to record both audio and video. Immediately after the interview, all video data is deleted. Nevertheless, the student may turn off his/her camera if preferred.
  - The student will receive the gift card for compensation after all interviews have been conducted.
  - Ask at which e-mail address the student prefers to receive the gift card.
- Run through warm-up item (a relatively short statistics task) to help the student understand what it means to think aloud and get past any potential nervousness.

Main part: 60 minutes

- Start recording.
- Present the statistics items one by one.
  - If the student stops speaking, remind him/her to remember to think out loud.
  - If the student gets stuck on an item and can't proceed, continue with the next item.

Reflection: 20 minutes

- After the last item *or* after approximately 60 minutes, the interviewer goes through the items again to ask the student for any clarification of their answer if necessary. If a student was vague on an item they got correct, it is not assumed their reasoning was right – the interviewer will ask questions to be sure. The correct answer to each item is only provided to the student after the questions of the interviewer are answered.

3 Interview coding

This subsection provides detailed information about how the data were coded, how biases in coding were avoided, and how alternative interpretations have been considered and rejected. We

used an extensive coding scheme, which is summarized in Table S1, followed by a description of each coding element. In the analysis, we considered to take a subset of students conditional on coding categories and for example only use data from students who understood the items. However, even when students revealed misunderstanding of certain concepts, they generally used the relevant attributes in their reasoning. Furthermore, as the sample size is small, subsetting would lead to very sparse data. Therefore, we kept all students in the analysis. We present the extensive coding scheme here, because it can be helpful for other researchers to examine similar topics.

Table S1: Coding scheme for student think-aloud study

Element	Coding categories
Provided response	Letter of choice (a, b, c, d, e, f); true (T); false (F); valid (V); invalid (I)
Correctness of response	Correct (1); incorrect (0)
Justification of response	Correct statistical reasoning (C); incorrect statistical reasoning (I); elimination (E); random guess (G); other (O)
Understanding of item	Yes (Y); some confusion (C); no (N)
Unknown statistical concepts	Yes, indicated by student (S); yes, revealed in reasoning (R); no (N)
Learning objective use	Evidence of use (1); no evidence of use (0)

Provided response Students' answers to each item were coded (letter of choice; true; false; valid; invalid). If a student selected a response but revised their answer later, the last given answer is coded.

Correctness of response For each answer, it was coded whether the answer was correct or incorrect.

Justification of response We coded what justification was shown for the provided answer: correct statistical reasoning, incorrect statistical reasoning, elimination, random guess, or other. The following guidelines are used:

- If a student's reasoning is partly correct and partly incorrect, this was coded as incorrect statistical reasoning. For example, one student was interpreting the meaning of a negative standard deviation (item *ARTIST_sc_MS_05*), and reasoned as follows: *"There is variability in scores, so not everybody scores the same, otherwise the standard deviation would be zero. [..] If the standard deviation is negative, then they probably have a lot of negative scores."* Whereas the first part of the reasoning is correct and the student shows understanding of the standard deviation as a measure of variability, the second part is incorrect; a negative standard deviation cannot be obtained. This was coded as incorrect statistical reasoning.

- If a student relied only partly on (in)correct statistical reasoning and in addition relied on elimination or guessing, this was coded as (in)correct statistical reasoning, since we are interested in identifying any statistical reasoning processes that students base their answers on.
- If a student relied partly on elimination and guessed their answer based on the remaining choice options, this was coded as a guess.

Understanding of item It was coded whether or not students showed understanding of the item or experienced some confusion, where understanding does *not* refer to the understanding of statistical concepts that were assessed but to e.g. the phrasing of the item in terms of ambiguity or unknown terms and examples.

Unknown statistical concepts It was also coded whether unknown statistical concepts were involved in the item, either if the student indicated this or if the student in their reasoning revealed misinterpretation or misunderstanding of a statistical concept that was involved in the item or that was required as prerequisite knowledge to answer the item. This coding element is relevant to verify whether students miss prerequisite knowledge that is required to engage in the reasoning processed necessary to answer the item. If student is familiar with the concept, but applies it incorrectly to the item, this is coded as *not* including unknown statistical concepts.

Learning objective use For the items with (partly) justified answers, it was coded for each learning objective whether students showed (complete or incomplete) evidence of using it or did not show evidence of using it. The following guidelines are used:

- Students may use different terminology than used in the learning objectives, but if the reasoning shows evidence of understanding of the principles that are addressed by the learning objective, this is coded as evidence of use.
- Evidence of use does not imply evidence of *correct* use. For example, learning objective 19 is formulated as: "Ability to make informal inferences about sample means based on measures of sampling variability.". Several students attempted to make such inferences about sample means, but used the standard deviation rather than the standard error as a measure of variability. This was still coded as use of learning objective 19, despite the incorrect application.
- If a learning objective is only partly used, this is coded as evidence of use. For example, learning objective 25 is formulated as "Understanding of how increasing the sample size increases power by reducing the standard error". Several items only require knowledge of how sample size influences power, but not the role of the standard error herein. If a students reasons about this without recognizing the role of the standard error, this is coded as evidence of use of learning objective 25.
- The student must show substantive use of a learning objective in their reasoning and not only mention concepts that the learning objective encompasses. This is mainly relevant for

the learning objectives of attribute A (Understanding center & spread). This attribute encompasses concepts that form the basis of many other statistical ideas, such as the mean and standard deviation. Only if students base their reasoning on a description or interpretation of these prerequisite concepts, this is coded as evidence of use of these learning objectives, but not if a concept is solely mentioned without elaborating on its meaning. For example, if a student indicates that the median can be seen in a boxplot, this is *not* coded as measuring learning objective 2 ("Ability to describe and interpret measures of center (mean, median mode).".

- If a student uses learning objectives to eliminate alternative choices, this is coded as evidence of use, even if the student uses solely other learning objectives to argue why the given choice is the correct answer.

III Assessment items

The assessment items were collected from several sources, namely the Statistical Reasoning Assessment (SRA; Garfield, 2003), the Statistics Concept Inventory (SCI; Allen, 2006), the Comprehensive Assessment of Outcomes in Statistics test (CAOS; delMas et al., 2007), and the Assessment Resource Tools for Improving Statistical Thinking (ARTIST; delMas et al., 2007). Minor adjustments were made to some items and the final items are presented here. The name of each item starts with the abbreviation of the source from which the original item was collected. The correct answers are indicated in italics.

Assessment 1: Samples and spread

Item ARTIST_sc_MS_05

A teacher gives a 15 item science test. For each item, a student receives one point for a correct answer; 0 points for no answer; and loses one point for an incorrect answer. Total test scores could range from –15 points to +15 points. The teacher computes the standard deviation of the test scores for the class to be –2.30.

What do we know?

a. The standard deviation was calculated incorrectly.
b. Most students received negative scores.
c. Most students scored below the mean.
d. None of the given options.

Item ARTIST_sc_MS_01

A class of 30 introductory statistics students took a 15 item quiz, with each item worth 1 point. The standard deviation for the resulting score distribution is 0.

You know that:

a. about half of the scores were above the mean.
b. an arithmetic error must have been made.
c. everyone correctly answered the same number of items.
d. the mean, median, and mode must all be 0.

Item ARTIST_sc_MC_05 & ARTIST_sc_MC_06

A college statistics class conducted a survey of how students spend their money. They gathered data from a large random sample of college students who estimated how much money they typically spent each week in different categories (e.g., food, entertainment, etc.). The following statistics were calculated for money spent weekly on food: mean = €31.52; median = €30.00; interquartile range = €34.00; standard deviation = €21.60; range = €132.50.

A student states that the median food cost tells you that a majority of students in this sample spend about €30 each week on food. Is this correct or incorrect, and why?

a. Correct, the median is an average and that is what an average tells you.
b. Correct, €30 is representative of the data.
c. Incorrect, a majority of students spend more than €30.
d. Incorrect, the median tells you only that 50% of the sample spent less than €30 and 50% of the sample spent more.

The students determined that a mistake had been made and a value entered as 138 should have been entered as 38. They recalculate all of the statistics. Which of the following would be true?

a. The value of the median decreases, the value of the mean stays the same.
b. The values of the median and mean both decrease.
c. The value of the median stays the same, the value of the mean decreases.

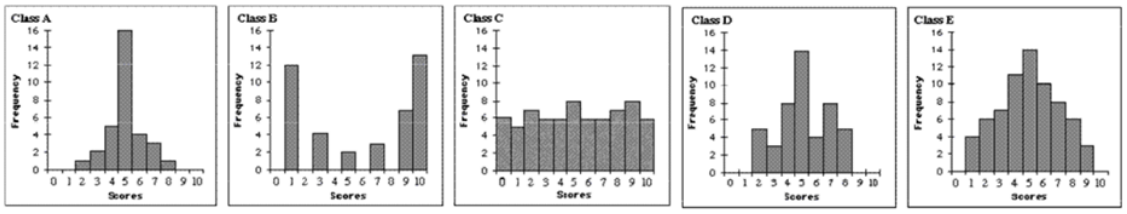
Item ARTIST_db_MS_Q0490

A mathematics test has been administered to 25 students in a high school class. The mean test score is 6.5 points and the standard deviation is 1.2. However, it turns out that the teacher miscalculated the scores and everybody receives exactly 1 point bonus. This results in an mean score of 7.5. How does this affect the size of the standard deviation?

a. The higher the mean, the higher the standard deviation. Therefore, the standard deviation increases.
b. The size of the standard deviation is not affected by this.
c. You need the mean to calculate the standard deviation. Therefore, the standard deviation changes as the mean changes, but we cannot tell whether it increases or decreases.

Item CAOS_14 & CAOS_15

Five histograms are presented below. Each histogram displays test scores on a scale of 0 to 10 for one of five different statistics classes.



Which of the classes would you expect to have the lowest standard deviation, and why?

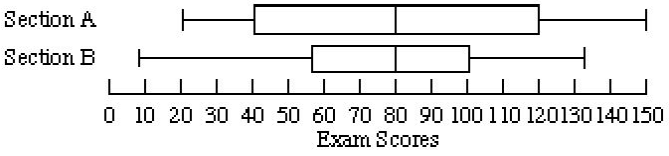
a. Class A, because it has the most values close to the mean.
b. Class B, because it has the smallest number of distinct scores.
c. Class C, because there is no change in scores.
d. Class A and Class D, because they both have the smallest range.
e. Class E, because it looks the most normal.

Which of the classes would you expect to have the highest standard deviation, and why?

a. Class A, because it has the largest difference between the heights of the bars.
b. Class B, because more of its scores are far from the mean.
c. Class C, because it has the largest number of different scores.
d. Class D, because the distribution is very bumpy and irregular.
e. Class E, because it has a large range and looks normal.

Item CAOS_08, CAOS_09 & CAOS_10

The two boxplots below display final exam scores for all students in two different sections of the same course.



Which section would you expect to have a greater standard deviation in exam scores?

a. Section A.
b. Section B.
c. Both sections are about equal.
d. It is impossible to tell.

Which data set has a greater percentage of students with scores at or below 30?

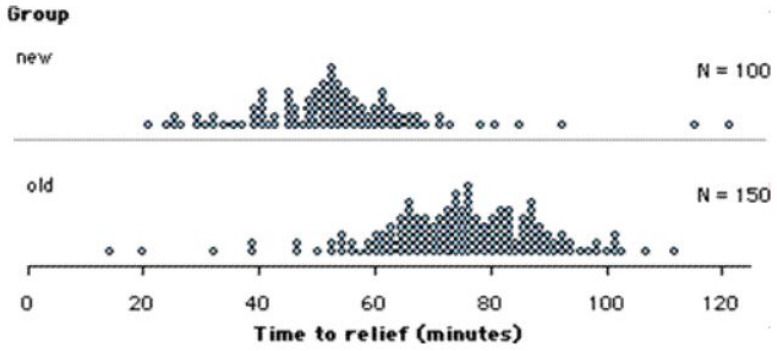
a. Section A.
b. Section B.
c. Both sections are about equal.
d. It is impossible to tell.

Which section has a greater percentage of students with scores at or above 80?

a. Section A.
b. Section B.
c. Both sections are about equal.
d. It is impossible to tell.

Item CAOS_12 & CAOS_13

A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches.100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown below.



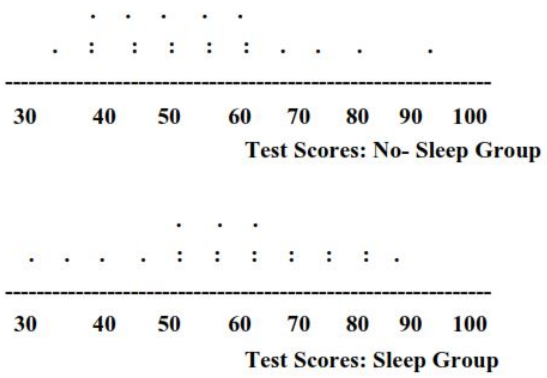
The questions below present statements made by two different statistics students. For each statement, indicate whether you think the student's conclusion is valid.

The average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.
Valid.

I would not conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.
Invalid.

Item SRA_015

Forty college students participated in a study of the effect of sleep on test scores. 20 of the students stayed up all night studying the night before the test (no-sleep group). The other 20 students (the control group) went to bed by 23:00 on the evening before the test. The test scores for each group are shown in the graphs below. Each dot on the graph represents a particular student's score. For example, the two dots above the 80 in the bottom graph indicate that two students in the sleep group scored 80 on the test.

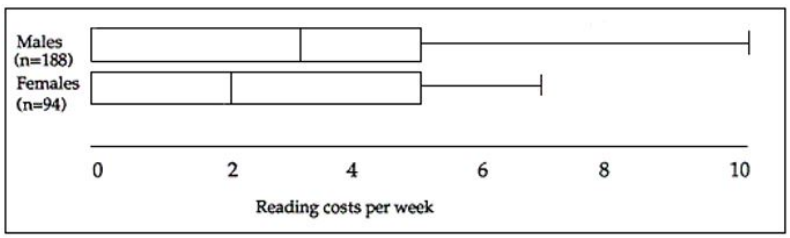


Examine the two graphs carefully. Then choose from the 6 possible conclusions listed below the one you *most* agree with.

a. The no-sleep group did substantially better because none of these students scored below 35 and the highest score was achieved by a student in this group.
b. The no-sleep group did substantially better because its average appears to be a little higher than the average of the sleep group.
c. There is no substantial difference between the two groups because there is considerable overlap in the scores of the two groups.
d. There is no substantial difference between the two groups because the difference between their averages is small compared to the amount of variation in the scores.
e. The sleep group did substantially better because more students in this group scored 80 or above.
f. The sleep group did substantially better because its average appears to be a little higher than the average of the no-sleep group.

Item ARTIST_db_CG_Q0840

Stephen wants to investigate differences in spending habits of males and females. He compares the amounts spent per week on reading materials by males and females in a random sample of college students by generating the following plots.



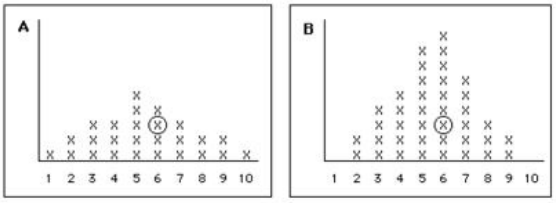
Stephen believes that his plots show that males and females tend to spend substantially different amounts of money on reading materials. Is this correct or incorrect, and why?

a. Correct, males tend to spend more money than females.
b. Correct, the median spending money for males is higher than the median for females.

c. Incorrect, the difference in spending is due to the fact that more males than females answered the survey.
d. Incorrect, the difference in medians is not large compared to the variability (IQR) of the data.
e. Incorrect, you can't draw conclusions about differences because there are more males than females in the sample.

Item ARTIST_sc_SV_01

Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights of random samples of 3 pebbles each, with the mean weights rounded to the nearest gram. One value is circled in each distribution.



Is there a difference between what is represented by the X circled in A and the X circled in B? Please select the best answer from the list below.

a. No, in both Figure A and Figure B, the X represents one pebble that weights 6 grams.
b. Yes, Figure A has a larger range of values than Figure B.
c. Yes, the X in Figure A is the weight for a single pebble, while the X in Figure B represents the average weight of 3 pebbles.

Item CAOS_017

Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 35% yellow candies. Five students each take – one at a time – a random sample of 20 candies and record the percentage of yellow candies in their sample. After each student records the percentage in their sample, they put the candies back in the barrel before the next student takes a sample. Which sequence below is the most plausible for the percent of yellow candies obtained in these five samples?

a. 30%, 35%, 15%, 40%, 50%.
b. 35%, 35%, 35%, 35%, 35%.
c. 5%, 60%, 10%, 50%, 95%.
d. Any of the given options.

Item ARTIST_sc_SV_03

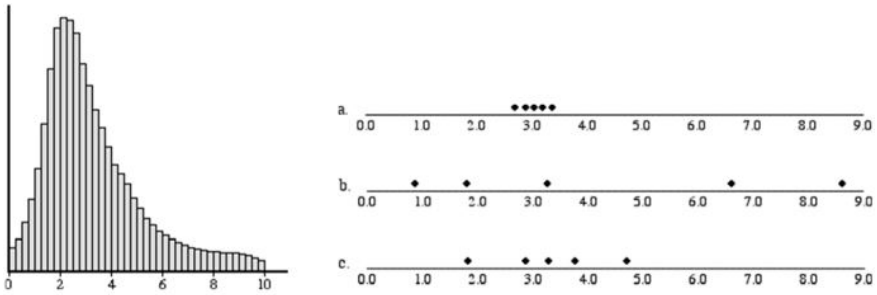
Suppose half of all newborns are girls and half are boys. Hospital A, a large city hospital, records an average of 50 births a day. Hospital B, a small, rural hospital, records an average of 10 births a day. On a particular day, which hospital is less likely to record 80% or more female births?

a. Hospital A (with 50 births a day), because the more births you see, the closer the proportions will be to .5.
b. Hospital B (with 10 births a day), because with fewer births there will be less variability.
c. The two hospitals are equally likely to record such an event, because the probability of a boy does not depend on the number of births.

Item ARTIST_sc_SV_14

The distribution for a population of measurements is presented below on the left. The mean is 3.2 and the standard deviation is 2. Suppose that five students each take a sample of ten values from the population and each student calculates the sample mean for his or her ten data values. The students draw a dotplot of their five sample means on

the classroom board so that they can compare them. Below on the right three possible outcomes for this dotplot are presented.



Which of the dotplots do you think is the most plausible for the one they drew on the board?

a. Plot a, because the sample means will all be almost equal to the population mean.
b. Plot b, because the sample means can take on any value in the range of the population distribution.
c. Plot c, because the sample means will be clustered around the population mean with some variability.

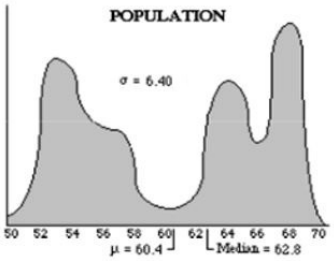
Item ARTIST_sc_SV_04

A random sample of 25 college statistics textbook prices is obtained and the mean price is computed. To determine the probability of finding a random sample with a more extreme mean than the one obtained from this random sample, you would need to refer to:

a. the population distribution of all college statistics textbook prices.
b. the distribution of prices for this sample of college statistics textbooks.
c. the sampling distribution of textbook prices for all samples of 25 textbooks from this population.

Item ARTIST_sc_SV_10, ARTIST_sc_SV_11 & ARTIST_sc_SV_09

A hypothetical distribution for a population of test scores is displayed below. The population has a mean of 60.4, a median of 62.8, and a standard deviation of 6.40.



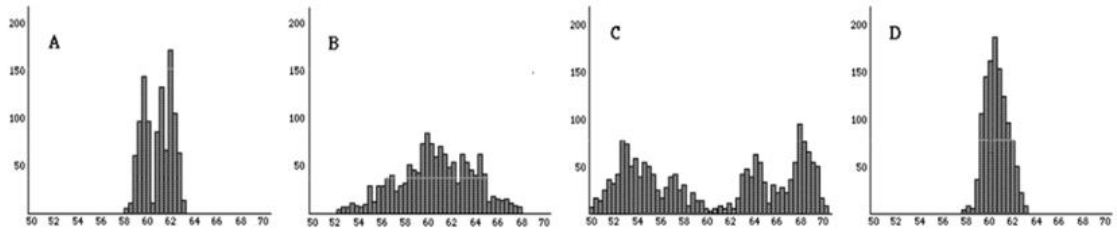
What do you expect for the shape of the sampling distribution for $n = 4$ (the distribution of sample means for all possible samples of size $n = 4$)?

a. Shaped more like a normal distribution than like the population distribution.
b. Shaped more like the population distribution than like a normal distribution.
c. Shaped like neither the population or the normal distribution.

What do you expect for the variability (spread) of the sampling distribution?

a. Same as the population.
b. Less variability than the population (a narrower distribution).
c. More variability than the population (a wider distribution).

Each of the other four graphs labeled A to D represent possible distributions of sample means for random samples drawn from the population. Which graph best represents a distribution of sample means for 1000 samples of size 4?



a. Graph A
b. Graph B
c. Graph C
d. Graph D

Item ARTIST_db_SS_Q0061A-D

Indicate for each of the following statements whether it is true or not true.

A. The mean of a sampling distribution of sample means is equal to the population mean divided by the square root of the sample size. *Not true.*
B. The larger the sample size, the more the sampling distribution of sample means resembles the shape of the population. *Not true.*
C. The mean of the sampling distribution of sample means for samples of size $n = 15$ will be the same as the mean of the sampling distribution for samples of size $n = 100$. *True.*
D. The larger the sample size, the more the sampling distribution of sample means will resemble a normal distribution. *True.*

Item ARTIST_sc_SV_05

Consider the distribution of average number of hours that college students spend sleeping each weeknight. This distribution is very skewed to the right, with a mean of 5 and a standard deviation of 1. A researcher plans to take a simple random sample of 18 college students. If we were to imagine that we could take all possible random samples of size 18 from the population of college students, the sampling distribution of average number of hours spent sleeping will have a shape that is:

a. Exactly normal.
b. Less skewed than the population.
c. Just like the population (i.e., very skewed to the right).
d. It is impossible to predict the shape of the sampling distribution. .

Item CAOS_016

A certain manufacturer claims that they produce 50% brown candies. Sam plans to buy a large family size bag of these candies and Kerry plans to buy a small fun size bag. Which bag is more likely to have more than 70% brown candies?

a. Sam, because there are more candies, so his bag can have more brown candies.
b. Sam, because there is more variability in the proportion of browns among larger samples.
c. Kerry, because there is more variability in the proportion of browns among smaller samples.
d. Kerry, because most small bags will have more than 50% brown candies.
e. Both have the same chance because they are both random samples. .

Item CAOS_032

It has been established that under normal environmental conditions, the average length of the population of largemouth bass in Silver Lake is 12.3 cm with a standard deviation of 3 cm. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group took a random sample of 100 largemouth bass from Silver Lake and found the mean of this sample to be 11.2 cm. Which of the following is the most appropriate statistical conclusion?

a. The researchers cannot conclude that the fish are smaller than what is normal because 11.2 cm lies within one standard deviation from the population mean (12.3 cm).
b. The researchers can conclude that the fish are smaller than what is normal because the sample mean should be almost identical to the population mean with a large sample of 100 fish.
c. The researchers can conclude that the fish are smaller than what is normal because the difference between 12.3 cm and 11.2 cm is much larger than the expected sampling error.

Item SCI_2004_20

The mean height of college men is 185 cm, with standard deviation 5 cm. The mean height of college women is 170 cm, with standard deviation 6 cm. You conduct an experiment at one university measuring the height of 100 male students and 100 female students. Which result would most surprise you?

a. One man with height 200 cm.
b. One woman with height 185 cm.
c. The average height of women in the sample is 175 cm.
d. The average height of men in the sample is 190 cm.

Item ARTIST_db_SS_Q1437

Other things being equal, as the sample size increases, the standard error of the mean:

a. increases
b. decreases
c. approaches the standard deviation of the population
d. approaches the population mean in numerical value.

Item ARTIST_db_SS_Q0614

A study was planned to examine the length of a certain species of fish on Gull Lake. Researchers took a random sample of 100 fish from this lake using a special net, and examined the results. Numerical summaries on lengths of the fish measured in this study are given.

Mean	25.018 cm
Median	25.295
Standard Deviation	4.1831
Range	20.73
Min	12.67
Max	33.4
N	78

The fish lengths in Lake Monster have a standard deviation that is twice as big as that on Gull Lake. Suppose you want to get an estimate of the mean fish length for Lake Monster with the same accuracy as for Gull Lake. Select the answer that best describes the sample size we need:

a. A larger sample is needed for Lake Monster than Gull Lake
b. A smaller sample is needed for Lake Monster than Gull Lake
c. About the same size sample
d. Not enough information to tell.

Assessment 2: NHST and confidence intervals

Item CAOS_40

The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?

a. The circuit is definitely not good and needs to be repaired.
b. The electrician decides that the circuit is defective, but it could be good.
c. The circuit is definitely good and does not need to be repaired.
d. The circuit is most likely good, but it could be defective.

Item ARTIST_sc_TS_01

The makers of Mini-Oats cereal have an automated packaging machine that is set to fill boxes with 650 grams of cereal. At various times in the packaging process, a random sample of 100 boxes is taken to see if the machine is filling the boxes with an average of 650 grams of cereal. Which of the following is a statement of the null hypothesis being tested?

a. The machine is filling the boxes with the proper amount of cereal.
b. The machine is not filling the boxes with the proper amount of cereal.
c. The machine is not putting enough cereal in the boxes.

Item ARTIST_db_TSG_Q1182

Which of the following is true?

a. It is impossible to prove the null hypothesis.
b. It is always possible to prove the null hypothesis.
c. It is possible to prove the null hypothesis under certain conditions.

Item CAOS_24

Herbicides can be used to control aquatic weeds, but can also unwantedly harm fish, which is indicated by higher levels of certain enzymes. A researcher in environmental science is conducting a study to investigate the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either a treatment or a control group. The treatment group is exposed to the herbicide and the control group is not. The fish in the treatment group showed higher levels of the indicator enzyme.

Suppose a test of significance was correctly conducted and showed a statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the researcher draw from these results?

a. The sample size is too small to draw a valid conclusion.
b. It is proven that the herbicide causes higher levels of the enzyme.
c. There is evidence that the herbicide causes higher levels of the enzyme.

Item ARTIST_db_TSG_Q1392

In a test to compare two populations means, which of the following alpha levels requires the least difference between the sample means in order to allow rejection of $H_0 : \mu_1 = \mu_2$? Assume all other factors are equal.

a. $\alpha = .11$
b. $\alpha = .06$

c. $\alpha = .04$
d. $\alpha = .007$
e. $\alpha = .003$

Item CAOS_25 & CAOS_26
A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with Macular Degeneration. The article gives a p-value of .04 in the analysis section. The questions below present three different interpretations of this p-value. Indicate if each interpretation is valid or invalid.

The p-value is the probability of getting results as extreme as or more extreme than the ones in this study if the drug is actually not effective.
Valid.

The p-value is the probability that the drug is not effective.
Invalid.

Item ARTIST_sc_TS_04
A researcher compares men and women on 100 different variables using an independent t-test. He sets the level of significance to .05 and then carries out 100 independent t-tests (one for each variable) on these data. If, for each test, the null hypothesis actually is true, about how many 'statistically significant' results will be produced?

a. 0
b. 5
c. 10
d. None of the given options

Item ARTIST_sc_TS_10 It is reported that scores on a particular test of historical trivia given to high school students are approximately normally distributed with a mean of 85. Mrs. Rose believes that her 5 classes of high school seniors will score significantly better than the national average on this test. At the end of the semester, Mrs. Rose administers the historical trivia test to her students. The students score an average of 89 on this test. After conducting the appropriate statistical test (with $\alpha = .05$), Mrs. Rose finds that the p-value is .0025. Which of the following is the best interpretation of the p-value?

a. A p-value of .0025 provides strong evidence that Mrs. Rose's class outperformed high school students across the nation.
b. A p-value of .0025 indicates that there is a very small chance that Mrs. Rose's class outperformed high school students across the nation.
c. A p-value of .0025 provides evidence that Mrs. Rose is an exceptional teacher who was able to prepare her students well for this national test.
d. None of the given options.

Item ARTIST_sc_TS_07
A newspaper article claims that the average age for people who receive food stamps is 40 years. You believe that the average age is less than that. You take a random sample of 100 people who visit the food bank, and find their average age to be 39.2 years. You find that this is significantly lower than the age of 40 stated in the article ($p < .05$). What would be an appropriate interpretation of this result?

a. The statistically significant result indicates that the majority of people who receive food stamps is younger than 40.
b. Although the result is statistically significant, the difference in age is not of practical importance.

c. An error must have been made. This difference is too small to be statistically significant.

Item ARTIST_sc_TS_09

A researcher conducts an experiment on human memory and recruits 15 people to participate in her study. She performs the experiment and analyzes the results. She obtains a p-value of .17. Which of the following is a reasonable interpretation of her results?

a. This proves that her experimental treatment has no effect on memory.
b. There could be a treatment effect, but the sample size was too small to detect it.
c. She should reject the null hypothesis.
d. There is evidence of a small effect on memory by her experimental treatment.

Item SCI_2004_22

You perform the same two significance tests on large samples from the same population. The two samples have the same mean and the same standard deviation. The first test results in a p-value of 0.01; the second in a p-value of 0.02. Which test has a larger sample size?

a. The first test.
b. The second test.
c. The sample sizes are equal in both tests.
d. We don't have enough information to determine which sample size is larger.

Item ARTIST_db_TSG_Q1007

A researcher conducts a study with a small sample size (i.e., n is small). Using a smaller sample size increases the risk of making the following error:

a. reject H_0 when H_0 is true.
b. reject H_0 when H_0 is false.
c. fail to reject H_0 when H_0 is true.
d. fail to reject H_0 when H_0 is false.
e. all of the above occur frequently in this situation.

Item ARTIST_sc_CI_05

Justin and Hayley conducted a mission to a new planet, Planet X, to study arm length. They took a random sample of 100 Planet X residents and calculated a 95% confidence interval for the mean arm length. What does a 95% confidence interval for arm length tell us in this case? Select the best answer:

a. I am 95% confident that this interval includes the sample mean arm length.
b. I am confident that most (95%) of all Planet X residents will have an arm length within this interval.
c. I am 95% confident that most Planet X residents will have arm lengths within this interval.
d. I am 95% confident that this interval includes the population mean arm length.

Item ARTIST_sc_CI_02

A 95% confidence interval is computed to estimate the mean household income for a city. Which of the following values will definitely be within the limits of this confidence interval?

a. The population mean
b. The sample mean
c. The standard deviation of the sample mean
d. None of the given options

Item ARTIST_sc_CI_01

Two different samples will be taken from the same population of test scores where the population mean and standard deviation are unknown. The first sample will have 25 data values, and the second sample will have 64 data values. A 95% confidence interval will be constructed for each sample to estimate the population mean. Which confidence interval would you expect to have greater precision (a smaller width) for estimating the population mean?

a. I expect the confidence interval based on the sample of 64 data values to be more precise.
b. I expect both confidence intervals to have the same precision.
c. I expect the confidence interval based on the sample of 25 data values to be more precise.

Item ARTIST_sc_CI_07

A pollster took a random sample of 100 students from a large university and computed a 95% confidence interval to estimate the percentage of students who were planning to vote in the upcoming election. The pollster felt that the confidence interval was too wide to provide a precise estimate of the population parameter. What could the pollster have done to produce a narrower confidence interval that would produce a more precise estimate of the percentage of all university students who plan to vote in the upcoming election?

a. Increase the sample size to 150.
b. Increase the confidence level to 99%.
c. Increase the sample size to 150 *and* increase the confidence level to 99%.
d. None of the given options.

Item ARTIST_sc_CI_06

Suppose that a random sample of 41 state college students is asked to measure the length of their right foot in centimeters. A 95% confidence interval for the mean foot length for students at this university turns out to be $(21.71, 25.09)$. If instead a 90% confidence interval was calculated, how would it differ from the 95% confidence interval?

a. The 90% confidence interval would be narrower.
b. The 90% confidence interval would be wider.
c. The 90% confidence interval would be the same as the 95% confidence interval.

Item ARTIST_sc_CI_10

Suppose two researchers want to estimate the proportion of college students who favor abolishing the penny. They both want to have about the same margin of error to estimate this proportion. However, Researcher 1 wants to estimate with 99% confidence and Researcher 2 wants to estimate with 95% confidence.

Which researcher would need more students for her study in order to obtain the desired margin of error (i.e., the desired width of the confidence interval)?

a. Researcher 1.
b. Researcher 2.
c. Both researchers would need the same number of subjects.
d. It is impossible to obtain the same margin of error with the two different confidence levels.

Item ARTIST_db_CIOSM_Q1394

Researchers ask a random sample of apartment dwellers in a large city their ideal air temperatures. They find the sample mean is 22 degrees Celsius. Using a two-tailed test, they reject $H_0 : \mu = 18$ at the 5% significance level. Which of the following could be a 95% confidence interval for μ, the average ideal temperature for all apartment dwellers in the city?

a. 21 – 27

b. 20 – 24
c. 18 – 26
d. 16 – 28
e. 16 – 20
f. Not enough information is given to answer this question

Item ARTIST_db_CIOSM_Q1387

A candy manufacturer produces larges bags with candies. To get an idea of the number of candies in those bags, a researcher takes a random sample of candy bags. The 95% confidence interval for the average number of candies per bag is $[40, 48]$. The researcher wants to use this confidence interval to test the null hypothesis $H_0 : \mu = 38$.

Based on this information, do we reject H_0 or not? At which α level do we draw conclusions about H_0?

a. We do not reject H_0 at α = .05
b. We do not reject H_0 at α = .10
c. We do not reject H_0 atα = .95
d. We reject H_0 at α = .05
e. We reject H_0 at α = .95

IV Final model results

The final model results are included the file `FinalModelResults.RData`. With these results, the mastery status of new students can be determined directly after they complete an assessment based on their automatically scored item responses. For each assessment, the following estimated parameters are included:

- Item response probabilities π_{ic}: the conditional probability that a respondent with attribute profile $\boldsymbol{\alpha}_c$ responds correctly to item i. These are contained in the R objects `pi1` and `pi2` for assessment 1 and 2 respectively.
- Structural parameters ν_c: the base-rate proportions of students with attribute profile c. These are contained in the R objects `nu1` and `nu2` for assessment 1 and 2 respectively.

These parameters can be used to estimate the posterior probability that a student has a certain attribute profile $\boldsymbol{\alpha}_c$ ($c = 1, ..., C$) given their item responses using Bayes' theorem:

$$\alpha_{rc} = P(\boldsymbol{\alpha}_c \mid \boldsymbol{x}_r) = \frac{P(\boldsymbol{x}_r \mid \boldsymbol{\alpha}_c)P(\boldsymbol{\alpha}_c)}{P(\boldsymbol{x}_r)} = \frac{\nu_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1-\pi_{ic})^{(1-x_{ir})}}{\sum_{c=1}^{C} \nu_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1-\pi_{ic})^{(1-x_{ir})}} \tag{1}$$

Here, $\boldsymbol{x}_r$ is the vector of response data from respondent r, x_{ir} is the observed response of respondent r to item i ($i = 1, ..., I$). To obtain marginal attribute information, the expected value for each attribute is computed across all attribute profiles:

$$P(\alpha_{ra} = 1 \mid \boldsymbol{\alpha}_r) = \sum_{c=1}^{C} \alpha_{ca} \alpha_{rc} \tag{2}$$

Here, α_{ca} is the attribute mastery indicator for attribute a in attribute profile c. For more details about the estimation of attribute profiles, we refer to Rupp, Templin, and Henson (2010, Ch. 10).

References

Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics* (PhD dissertation). University of Oklahoma.

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Dordrecht: Springer. doi: 10.1007/1-4020-2278-6_7

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. e., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology*, *25*(4), 527–547. doi: 10.1080/0020739940250406

Batanero, C., Tauber, L. M., & Sánchez, V. (2004). Students' reasoning about the normal distribution. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 257–276). Dordrecht: Springer. doi: 10.1007/1-4020-2278-6_11

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389. doi: 10.1037/1082-989X.10.4.389

Bower, K. M. (2003). Some misconceptions about the normal distribution. In *American Society for Quality: Six Sigma Forum.*

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

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

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170. doi: 10.1037/0003-066X.60.2.170

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*(4), 299–311. doi: 10.1207/s15328031us0304_5

delMas, R., Garfield, J. B., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55–82.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75–98. doi: 10.1177/0959354395051004

Fidler, F. (2006). Should psychology abandon p-values and teach CIs instead? Evidence-based reforms in statistics education. In *Proceedings of the 7th International Conference on Teaching Statistics.* Salvador, Brazil.

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. doi: 10.1027/0044-3409.217.1.27

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

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22–38.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Science & Business Media. doi: 10.1007/978-1-4020-8383-9

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). Learning to reason about distribution. In *Developing students' statistical reasoning* (pp. 165–186). Springer. doi: 10.1007/978-1-4020-8383-9_8

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): what do the textbooks say? *The Journal of Experimental Education*, *71*(1), 83–92. doi: 10.1097/00004583-200102000-00021

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

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. doi: 10.3758/s13423-013-0572-3

Kalinowski, P., et al. (2010). Identifying misconceptions about confidence intervals. In *Proceedings of the 8th International Conference on Teaching Statistics* (Vol. 50). Ljubljana, Slovenia.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. *Research on the role of technology in teaching and learning statistics*, 151–167.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In *Proceedings of the 6th International Conference on Teaching Statistics.* Cape Town, South Africa.

Makar, K., & Confrey, J. (2002). Comparing two distributions: Investigating secondary teachers' statistical thinking. In *Proceedings of the 6th International Conference on Teaching Statistics.* Cape Town, South Africa.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, *6*, 223. doi: 10.3389/fpsyg.2015.00223

Reichardt, C. S., & Gollob, H. F. (2016). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 231–254). Routledge. doi: 10.4324/9781315629049

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

Schuyten, G. (1991). Statistical thinking in psychology and education. In D. Vere-Jones (Ed.), *Proceedings of the 3rd International Conference on Teaching Statistics: Vol. 2. Teaching Statistics Beyond School Level* (pp. 486–490). Dunedin, New Zealand: ISI Publications in Statistical Education.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105. doi: 10.1037/h0031322

Vallecillos, A. (2000). Understanding of the logic of hypothesis testing amongst university students. *Journal für Mathematik-Didaktik*, *21*(2), 101–123. doi: 10.1007/BF03338912

Vallecillos, A., & Batanero, C. (1997). Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios [Activated concepts in statistical hypothesis testing and their understanding by university students]. *Recherches en Didactique des Mathématiques (Revue)*, *17*(1), 29–48.

Williams, A. M. (1998). Students' understanding of the significance level concept. In *Proceedings of the 5th International Conference on Teaching Statistics.* Singapore.