EXAMINING THE RELATIONSHIP BETWEEN STATISTICAL LITERACY AND STATISTICAL REASONING

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Anelise Guimaraes Sabbag

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

Joan Garfield, Adviser
Andrew Zieffler, Co-adviser

June 2016

Acknowledgements

This dissertation is finished thanks to God, my advisers, my family, and my friends. I am immensely grateful for God's and Jesus' presence in my life throughout my studies giving me strength to finish all that They have laid before me. I am also immensely grateful for the amazing peace and great comfort in times of trouble that the Holy Spirit provided to me.

I also want to thank my advisers (Joan Garfield and Andy Zieffler) and other professors and scholars. I thank Joan for her guidance, encouragement and for helping me to become a statistical education researcher. Andy was with me since my first year at the University of Minnesota. He challenged me throughout the whole way! It was because of his willingness to advise me that I have reached this point in my education and for that I am so grateful. I thank Professor Mark Davison who was the great source of knowledge about measurement that helped me to finish this dissertation. He was not my adviser but he surely acted like he was. I also want to thank Bob delMas, Jane Watson, Maxine Pfannkuch, Rob Gould, Laura Le, and Elizabeth Fry for thoughtful feedback supporting my dissertation research.

I am also eternally grateful for one of the most important persons in my life: my father, Pedro. He gave me the opportunity to come to the US to study and he supported me throughout these five years. Without my father I would not have obtained a PhD degree.

My mom Dulce, my brother Mauricio and his wife Crislene, my grandpa Renato, and my friend Nicola were also the biggest source of encouragement through the hard times of graduate school.

Finally, I want to thank Kim Strain for giving me the advice and tools to help improve my writing. And I also thank Ethan, Martin, Kory, Yadira, and all other friends for great insight and feedback.

Dedication

"For from Him and through Him and for Him are all things. To Him be the glory forever!"
Romans 11:36

Abstract

Statistical literacy and statistical reasoning have been considered by the statistics education community as important learning goals to be developed in introductory statistics courses (Garfield & Ben-Zvi, 2008). Many statistics educators and scholars have tried to define these learning goals (e.g., Gal, 2002; Watson & Callingham, 2003; Garfield, 2002; Garfield & Chance, 2000; Wild & Pfannkuch, 1999). However, there is a lack of agreement regarding these definitions and the relationship between statistical literacy and statistical reasoning. In addition, there are assumptions in the statistics education literature of an overlap between statistical literacy and statistical reasoning (e.g., Rumsey, 2002; Budgett & Pfannkuch, 2007, and delMas, 2002) and of a hierarchy between these learning goals (Garfield & Ben-Zvi, 2008 and delMas, 2002). Empirical evidence is needed to support these assumptions.

The purpose of this study was to investigate how statistical literacy and statistical reasoning are related. Specifically, this research aimed to verify if these two learning goals are distinct or if they overlap. The three research questions addressed by this study were (1) what measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony? (2) what measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction? (3) what measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning?

To answer the three research questions, the REALI instrument was developed to concurrently measure statistical literacy and statistical reasoning. This instrument is composed of 40 items with 20 items measuring statistical literacy and 20 item measuring statistical reasoning. The items in this instrument assess eight areas of learning: (1) representations of data, (2) measures of center, (3) measures of variability, (4) study design, (5) hypothesis testing and p-values, (6)

confidence intervals, (7) bivariate data, and (8) probability. During the development process, several types of validity evidence were gathered to support the intended inferences and uses of the REALI's scores (and subscores): expert reviews, response process interviews with students, a pilot test, a field test, reliability, and a psychometric analysis.

Data from the field test were analyzed under the classical test theory (CTT) and item response theory frameworks (IRT). Five IRT models were fitted to the data: A Unidimensional Model, three bi-dimensional models (Uncorrelated Model, Correlated Model, and Cross-loading Model), and a bi-factor model (Bi-factor Model A). The main difference between these models was whether or not the model allowed statistical literacy and statistical reasoning dimensions to correlate and if a third dimension (statistical knowledge) was included in the model. These models were compared at the item- and model-level and the best fitting models were used to evaluate the relationships between the statistical literacy subscore and the statistical reasoning subscore. Evidence was found that the statistical literacy and reasoning subscores from the Cross-loading Model could be measured reliably and distinctly. In addition, statistical evidence also supported that reporting both statistical literacy and statistical reasoning subscores provided more distinct information than only reporting a unique score for each student. Such findings bring valuable information to the field of statistics education and can be used to guide instruction in introductory statistics courses. In addition, the REALI instrument is a tool that can be used by researchers and instructors to investigate students' understanding of statistical concepts or to evaluate new curricula.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The field of statistics education has been going through many changes, one of which is a change in the learning goals in introductory statistics courses. This change involves a shift from computation and procedures to the goals of helping students develop statistical literacy, statistical reasoning, and statistical thinking (Garfield & Ben-Zvi, 2008).

Many statistics educators have tried to define and describe statistical literacy, reasoning, and thinking (e.g., Gal, 2002; Budgett & Pfannkuch, 2007; Rumsey, 2002; Watson & Callingham, 2003; Garfield, 2002; Garfield & Chance, 2000; Jones et al., 2004; Chance, 2002; Wild & Pfannkuch, 1999; Pfannkuch & Wild, 2000, 2003, 2004; delMas, 2004). However, no consensus has been reached regarding the definitions of these terms. The lack of consistency in the different definitions supports the idea that these concepts are still evolving. In addition, there seems to be a great overlap in the definitions of these terms and assumptions of a hierarchy between and within these learning goals has been posed by some researchers (delMas, 2002; Chance, 2002; Garfield & Ben-Zvi, 2007 and 2008). However, no empirical study has been done to examine the relationship between these learning goals.

A few instruments have been developed to assess some of the desired learning goals for student learning introductory statistics. Currently, there are three published assessments that measure statistical reasoning (SRA – Garfield, 1991; CAOS – delMas et al., 2007, and GOALS – Sabbag & Zieffler, 2015) at the tertiary level and one assessment measuring statistical literacy (BLIS – Ziegler, 2014). There is no published instrument available to assess statistical thinking. Because of the lack of instruments available to measure students' statistical thinking, this study will focus solely on statistical literacy and reasoning.

Despite the evidence of a possible overlap between the outcomes of statistical literacy and reasoning, there is no assessment measuring these learning goals concurrently. The assessments of statistical reasoning and statistical literacy that are currently available were developed independently without taking into account the possible overlap between these learning goals. However, if the overlap between statistical literacy and reasoning is so large that it is not possible to differentiate them, then the idea of having two separate learning goals (and two different assessments) should be re-evaluated. To investigate the overlap between statistical literacy and statistical reasoning, it is necessary to develop one instrument composed of some items measuring statistical literacy and other items measuring statistical reasoning. In this way it would be possible to obtain two subscores (one for each learning goal) and use measurement analysis to explore if statistical literacy and statistical reasoning could be measured reliably and distinctly or if these two learning goals are actually so similar that they cannot be distinguished. Thus having an instrument that measures statistical literacy and statistical reasoning at the same time, would help to clarify the structure of the relationship between these outcomes.

1.1 Description of the Study

This goal of this research study was to examine the relationship between statistical literacy and statistical reasoning by developing an assessment that concurrently measures these two learning goals. Three research questions are posed in this study: (1) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony? (2) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction? (3) What measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning?

1.2 The REALI Instrument

In this study a new assessment instrument was developed: the Reasoning and Literacy instrument (REALI). This instrument is composed of 40 items and was developed to concurrently measure statistical literacy and statistical reasoning in introductory statistics courses at the secondary level. Half of the items (20) in the instrument measures statistical literacy and the other half measures statistical reasoning. In addition, this instrument assesses eight areas of learning: (1) representations of data, (2) measures of center, (3) measures of variability, (4) study design, (5) hypothesis testing and *p*-values, (6) confidence intervals, (7) bivariate data, and (8) probability.

As suggested by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), evidence of validity was collected throughout the development process (e.g., expert reviews, response process interview with students, pilot and field test of the assessment, and psychometric analysis). The development process started with the creation of working definitions of statistical literacy and statistical reasoning based on well-known and widely used definitions available in the literature. The next stage in the process was the establishment of the basis of the assessment which was composed of items from a statistical literacy assessment (BLIS) and items from a statistical reasoning assessment (GOALS). However, some of these items were not aligned with the working definitions used in the study, therefore, some items were deleted and changed. In addition, some items were added to the instrument to meet the desired number of items (40 total items: 20 literacy and 20 reasoning) leading to the first draft of the assessment.

To provide content evidence of validity for the instrument, experts in the field of statistics education were asked to categorize the items, in the first draft of the instrument, as statistical literacy or statistical reasoning items. Experts were also asked to critique and make suggestions for the improvement of the items. Improvements were made to the items based on this expert feedback. In addition, the categorization of the items done by the experts was compared to the one done by

the researcher. Items that had a high level of disagreements regarding their categorization were further investigated by including such items in the think-aloud interview with students.

Think-aloud interviews were conducted with four University of Minnesota's students. The intent of these interviews was to gather students' response process validity evidence by examining how students were answering the items. During the interviews, students were given the items from the REALI assessment and asked to read each item out loud and say what they were thinking while answering the items. Students' responses were used to identify items that were not clear to the students and to improve such items.

The next stage in the development process involved a pilot study with 237 undergraduate students from the University of Minnesota and Augsburg College. The purpose of the pilot study was to investigate the psychometric properties of the items and identify items that were poorly behaving. Additional changes were done to the instrument, based on the quantitative analysis of the pilot data, and this led to the final version of the assessment which was used in the field test with 758 students from 16 universities and colleges around the United States. Students' answers from the field test were used to investigate the relationship between statistical literacy and statistical reasoning thus answering the three research questions posed by the study.

1.3 Overview of the Chapters

This chapter introduces the nature and importance of the study and also presents the research questions that will be addressed in the thesis. Chapter 2 provides a review of the literature in the field of statistics education related to the definitions of statistical literacy, statistical reasoning, and statistical thinking. In addition, this chapter describes the attempts to distinguish and assess these learning goals.

Chapter 3 presents a chronological description of the development stages of the new assessment including information about the blueprint, the process of adding and modifying items,

the expert review, the think aloud interviews, the pilot test, and the field test. This chapter also explains the procedures used for collecting and analyzing students' responses to answer the three research questions. Chapter 4 reports the changes and improvements made to the instrument during each of the development phases. This chapter also shows the results from the data analyses such as descriptive statistics, analysis of the structure of the instruments, and item response theory modelling. Finally, Chapter 5 provides answers to all research questions based on the results from the analyses. This chapter also describes the limitations of the study and the implications for future research. Following Chapter 5 are appendices including supporting documents such as the changes made to the items and copies of the different versions of the REALI instrument at various stages of its development.

Chapter 2

Review of the Literature

The purpose of this literature review is to better understand the constructs of statistical literacy, statistical reasoning, and statistical thinking. This chapter examines how statistical literacy, reasoning, and thinking have been described in the literature and how they can be assessed and distinguished. Research is also reported on two frameworks for classifying educational goals. Then, a summary and critique of the literature is offered followed by the problem statement and research questions.

2.1 Statistical Literacy

The definition of statistical literacy is still evolving and there has been much work to better explain and describe this term. This section reports what has been written about statistical literacy in the field of statistics education. Research on this topic will be reported in four different areas: statistical literacy in adults, college students, teachers and principals, and school students.

2.1.1 Statistical Literacy in Adults. One of the most cited definitions of statistical literacy is the one by Gal (2002). In his paper about the nature of adults' statistical literacy and its components, Gal (2002) proposed a model of statistical literacy with two components: knowledge and dispositions. The knowledge component is comprised of cognitive elements such as literacy skills, knowledge of statistics, mathematics and context, and critical questions. The dispositions component is comprised of critical stance and composed of beliefs and attitudes. In addition to proposing this model of statistical literacy, Gal also defined statistical literacy as the ability to comprehend, interpret, communicate, and critically evaluate statistical information.

Budgett and Pfannkuch (2007) built on Gal's (2002) definition of adults' statistical literacy. However, they added a reasoning piece to his definition. This piece was comprised of statistical

argumentation knowledge and everyday events knowledge (viewing daily events from a statistical perspective). Kaplan and Thorpe (2010) also focused on adults' statistical literacy and built on Gal's (2002) definition. The authors defined statistical literacy as the skills and knowledge adults need to be consumers of statistics. Based on five publications about statistical literacy (Cobb, 1992; ASA, 2005; Gal, 2002; Rumsey, 2002, and Utts, 2003), Kaplan and Thorpe identified five important content areas of statistical literacy: (1) data and experimental design, (2) probability, (3) variability, (4) descriptive statistics, and (5) conclusions and inferences. For each of these areas, Kaplan and Thorpe outlined what people would need to know and do.

2.1.2 Statistical Literacy in College Students. Statistical literacy has also been defined with a focus on college students, instead of adults. In her article about statistical literacy and how to develop it in introductory statistics courses, Rumsey (2002) reviewed many definitions of statistical literacy available in the literature. After this review, she decided to use the terms "statistical competence" and "statistical citizenship" instead of the term "statistical literacy". These two terms are related to two goals she aims to achieve in introductory statistics courses. According to Rumsey (2002), statistical competence refers to the knowledge which must be acquired before being able to reason and think statistically. Statistical Competence is therefore related to data awareness, data collection, basic statistical concepts, and basic interpretation and communication skills. Statistical citizenship was defined as a person's ability to operate in a data-driven society, in other words, being able to receive, critique, and make decisions based on statistical information (Rumsey, 2002).

Similar to the definition of statistical competence from Rumsey (2002), Garfield and delMas (2010) defined statistical literacy as knowing and applying statistical language and tools. A definition very similar to Garfield and delMas' definition of statistical literacy, was given by the

Guidelines for Assessment and Instruction in Statistics Education (GAISE; ASA, 2005). They defined statistical literacy as understanding key concepts, symbols, and language of statistics.

2.1.3 Statistical Literacy in K-12. A major area of interest has been the development and assessment of statistical literacy in school students. For instance, Watson and Callingham (2003) focused on grades 3 to 9 and Sharma, Doyle, Shandil, and Talakia'atu (2011) focused on grade 9 students. In their paper about the hierarchical nature of statistical literacy, Watson and Callingham (2003) used the structural cognitive model from Biggs and Collis (1982). They identified six hierarchical levels of statistical literacy: idiosyncratic, informal, inconsistent, consistent/non-critical, critical, and critical/mathematical. They also investigated the assumption of unidimensionality and they concluded that statistical literacy is a unidimensional construct. Sharma et al. (2011) also developed a statistical literacy framework based on Watson's research on statistical literacy. However, their framework was comprised of four components: informal/idiosyncratic, consistent noncritical, early critical, and advanced critical.

Much work has been done on statistical literacy; however, no clear consistency was observed among its many definitions. In addition, there is a lack of evidence regarding the construct of statistical literacy. For instance, Watson and Callingham (2003) reported many important findings about statistical literacy; however, care is needed when examining the conclusions of their study. They claimed that the assessment's items used in their study "covered the range of potential contributing elements to the construct of statistical literacy". However, the authors did not present any evidence regarding what the items were assessing and more importantly, if the items were indeed assessing statistical literacy. This could mean that the assessments they used could have measured various components of statistics instead statistical literacy itself**.** Therefore, content

validity evidence is missing in this study and consequently the conclusions about the construct of statistical literacy need to be viewed with great care.

Another area of research is statistical literacy in professionals. Chick and Pierce (2013) focused on how much statistical literacy do teachers and principals need to be able to read and interpret educational reports containing graphical information. Their ideas of statistical literacy were based on definitions from other researchers (e.g., Watson, 2006; Gal, 2002; and Curcio, 1987). However, these definitions did not address the topic of interest of the authors; therefore, they developed a new framework of statistical literacy. Because their study focused on teachers and principals, Chick and Pierce referred to *professional* statistical literacy and categorized it in three levels: (1) reading values, (2) comparing values, and (3) analyzing the data set. The first level, reading values, is related to making sense of the individual features of the graph. The second level of this framework, comparing values, contemplates recognizing multiple features of the graph to allow for comparison. Finally, the third level, analyzing the data set, involves recognition of the graph as a whole, use of statistical knowledge to interpret the information on the graph, and understanding how different data lead to different graphs. Despite the importance of statistical literacy, other goals are also desired for introductory statistics students.

2.2 Statistical Reasoning

Statistics educators have stated that statistical reasoning is also an important learning goal for students. This section presents the research on defining and describing statistical reasoning. Unlike statistical literacy, most of the definitions of statistical reasoning are primarily focused on college and school students. In addition, the majority of work in this area has focused on reasoning about specific statistical concepts.

2.2.1 Describing Statistical Reasoning. Like statistical literacy, statistical reasoning has been defined by many statistics educators and researchers. Garfield (2002) reviewed definitions of statistical reasoning from six researchers (Chervaney, Collier, Fienberg, Johnson, and Neter, 1977; Chervaney, Benson, and Iyer, 1980; Hawkins, Jolliffe, and Glickman, 1992; Nisbett, 1993; Sedlmeier, 1999; and Lovett, 2001) and she concluded no clear agreement has been reached regarding these definitions. In addition, she stated that more studies are needed to better understand students' statistical reasoning and how it can be developed in statistics courses.

In an attempt to clarify the learning goal of statistical reasoning, Garfield and Chance (2000) and Garfield (2002) defined statistical reasoning as reasoning with statistical concepts and understanding statistical information. According to the authors, statistical reasoning includes interpretations, representations and summarizing data. It also includes connecting statistical concepts from which further inferences can be drawn. Garfield and Dani Ben-Zvi (2008) expanded the previous definition, stating that statistical reasoning is "mental representations and connections that students have regarding statistical concepts" (page 34). Further efforts were made by Jones, Langrall, Mooney, and Thornton (2004) to better explain the construct of statistical reasoning. These authors reviewed three papers about models of development in statistical reasoning (Jones, Thornton, Langrall, Mooney, Perry, & Putt, 2000; Mooney, 2002; and Watson, Collis, Callingham, & Moritz, 1995) and stated that this construct was composed of hierarchical stages and cycles. Despite the definitions and explanations regarding the construct of statistical reasoning, the understanding of this concept is still evolving.

2.2.2 Reasoning about Statistical Concepts. Current work on statistical reasoning has been focused on individual concepts of statistics:

- Reasoning about statistical inference (e.g., Rossman, 2008; Lane-Getaz, 2013; Zieffler, Garfield, delMas, & Reading, 2008),
- Reasoning about data (e.g., Konold, Higgins, Russell, & Khalil, 2003; Konold & Higgins, 2003; Heaton and Mickelson, 2002),
- Reasoning about center (e.g., Jones, Thornton, Langrall, Mooney, Perry, and Putt, 2000; Mooney, 2002; Groth & Bergner, 2006; Bakker & Gravemeijer, 2006),
- Reasoning about statistical models and modelling, distribution, (e.g., Batanero, Tauber, and Sánchez, 2004; Pfannkuch & Reading, 2006; Wild, 2006),
- Reasoning about covariation (e.g., Zieffler & Garfield, 2009; Moritz, 2004; Zieffler, 2006),
- Reasoning about variability (e.g., Reading & Shaughnessy, 2004; Garfield & Ben-Zvi, 2005; Gould, 2004; Lehrer & Schauble, 2007),
- Reasoning about comparing groups (e.g., Konold & Higgins, 2003; Ben-Zvi, 2004b; Hammerman and Rubin, 2004; Bakker et al., 2004), and
- Reasoning about samples and sampling distribution (e.g., Watson, 2004; Bakker, 2004b; Chance, delMas, & Garfield, 2004).

Sometimes in the literature, the term statistical reasoning is used interchangeably with other terms, such as statistical thinking which is another important learning goal for introductory statistics courses.

2.3 Statistical Thinking

Statistical thinking is another important learning goal for statistics students. Similar to statistical literacy and reasoning, no agreement has been reached regarding the definition of

statistical thinking (Chance, 2002). This section presents the research on defining and describing statistical thinking.

One of the most cited definitions of statistical thinking is the one by Moore (1990). He defined it as a "general way of thinking in the realm of inquiry" and stated the five elements of statistical thinking as (1) the need for data about processes, (2) the design of data production with variation in mind, (3) the omnipresence of variation in processes, (4) quantification of variation, and (5) explanation of variation. This definition was the foundation for other definitions used in two educational reports: *Heading the Call for Change* from the Mathematical Association of America (Cobb, 1992) and *Guidelines for Assessment and Instruction in Statistics Education* from the American Statistics Association (ASA, 2005). Another early definition of statistical thinking came from the area of total quality. Snee (1990) defined statistical thinking in quality improvement as the awareness of the omnipresence of variability and the need to model and reduce variability.

After reviewing definitions of statistical thinking from leading statisticians, Mallows (1998) criticized these definitions stating that they did not emphasize or consider important the thought process regarding the importance of the observed data. The author then stated that statistical thinking most likely happens when variability and uncertainty are present and it relates problems that actually exist with quantitative data. He added that statistical thinking uses the data to accurately explain the considered issue or real world problem.

Empirical studies investigating the construct of statistical thinking were performed by Wild and Pfannkuch (1999), Pfannkuch and Wild (2000, 2003, 2004). After interviewing six statisticians and 16 statistics students about their reasoning process when solving statistical problems, Wild and Pfannkuch (1999) reported a framework for statistical thinking composed of four dimensions: (1) investigative cycle, (2) types of thinking, (3) interrogative cycle, and (4) dispositions.

- The first dimension (investigative cycle) is related to how people think and act when performing statistical analysis. This dimension is composed of five stages of investigation: Problem, Plan, Data, Analysis, and Conclusions.
- The second dimension (types of thinking) involves two types of thinking: general thinking (strategic, seeking explanations, modelling, and applying techniques) and thinking fundamental to statistical thinking (recognition of need for data, transnumeration, consideration of variation, reasoning with statistical models, and integrating the statistical a contextual).
- The third dimension (interrogative cycle), is a five stage thinking process that occurs when a person is solving a problem (generating, seeking, interpreting, criticizing, and judging).
- The final dimension (dispositions), is composed of eight thinking dispositions: (1) skepticism, (2) imagination, (3) curiosity and awareness, (4) openness, (5) propensity to seek deeper meaning, (6) being logical, (7) engagement, and (8) perseverance.

Despite some critiques (e.g., Moore, 1999; Breslow, 1999; Snee, 1999), this framework from Wild and Pfannkuch (1999) has been widely used in the field of statistics education (e.g., Sanchez & Blancarte, 2008; Makar & Confrey, 2002; Groth, 2005) and was the basis for other definitions of statistical thinking. For instance, Garfield and Ben-Zvi (2007, 2008) built on the definition from Wild and Pfannkuch (1999) and stated that statistical thinking

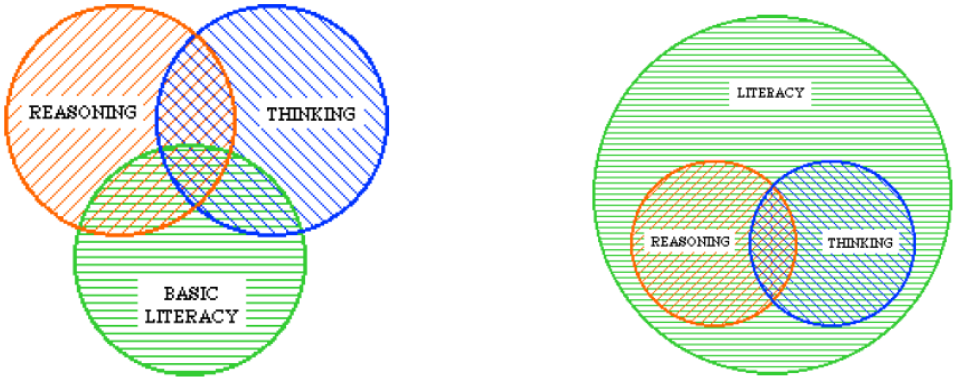
> …includes knowing how and why to use a particular method, measure, design or statistical model; deep understanding of the theories underlying statistical processes and methods; as well

> as understanding the constraints and limitations of statistics and statistical inference. Statistical thinking is also about understanding how statistical models are used to simulate random phenomena, understanding how data are produced to estimate probabilities, recognizing how, when, and why existing inferential tools can be used, and being able to understand and utilize the context of a problem to plan and evaluate investigations and to draw conclusions (Chance, 2002). Finally, we view statistical thinking as the normative use of statistical models, methods, and applications in considering or solving statistical problems. (p. 34)

Like statistical literacy and reasoning, the definition of statistical thinking is still evolving. Moreover, after examining the literature, the overlap among definitions of statistical reasoning, literacy, and thinking is evident. Therefore, an examination of how to distinguish these outcomes is reported in the next section.

2.4 Distinguishing Statistical Literacy, Reasoning and Thinking

The terms statistical literacy, reasoning, and thinking have been used interchangeably in the literature (Garfield, 2002; Chance, 2002; delMas, 2004). In an attempt to understand the possible connections between these learning goals, delMas (2002) reviewed definitions of statistical reasoning, literacy, and thinking from Garfield (2002), Rumsey (2002), and Chance (2002), respectively. Based on this review, he posed two models to represent the relationships between these constructs. In the first relationship, there is an overlap between statistical reasoning, literacy, and thinking (Figure 2.1a). However, each construct also has an independent part. In the second proposed relationship, statistical reasoning is still overlapping with statistical thinking; however, both constructs are presented as a subset of statistical literacy (Figure 2.1b). Therefore, there is no independent content among statistical literacy and reasoning and among statistical literacy and thinking.



(a) (b)

Figure 2.1. Models for statistical literacy, reasoning and thinking.

Instead of focusing on possible models of the relationship between statistical literacy, reasoning, and thinking, delMas (2002) argued that it was possible to differentiate among these learning goals when examining the nature of the task performed. For instance, students can demonstrate statistical reasoning when asked to provide an *explanation* or *justification* of statistical results (delMas 2004, 2002). delMas (2002) also presented different vocabulary that could be used to distinguish and assess statistical reasoning, literacy, and thinking. For instance, the author stated that tasks asking "why", "how", or "explain (the process)" could be considered statistical reasoning tasks and tasks asking "apply", "critique", "evaluate" or "generalize" could be considered statistical thinking tasks.

Chance (2002) also attempted to differentiate statistical thinking from statistical literacy and reasoning. Chance stated that a statistical thinker goes beyond what was taught in the statistics course and is able to understand the whole statistical process. In addition, statistical thinking involves moving beyond the discernment and explanation of statistical information (statistical literacy) and moving beyond concepts and skills (statistical reasoning). A statistical thinker performs spontaneous investigation, questioning and exploration of the problems and data.

Garfield and Ben-Zvi's (2007, 2008) view of statistical literacy, reasoning and thinking is aligned with the first model (Figure 2.1a) presented by delMas (2002). In addition, they also stated the existence of a hierarchy between them, with statistical literacy as the basis for the other two constructs and with statistical thinking requiring a higher order of thinking than statistical reasoning. To better understand this possible hierarchy, Garfield and Ben-Zvi (2008) compared statistical literacy, reasoning, and thinking to other hierarchical frameworks for the classification of students' learning.

2.5 Frameworks for Classification of Educational Goals

Bloom's taxonomy and Structural Observation of Learning Outcomes (SOLO) are two widely used taxonomies to classify students' understanding in hierarchical categories.

2.5.1 Bloom's Taxonomy. Bloom's taxonomy was designed "to be a classification of the student behaviors which represent the intended outcomes of the educational process" (Bloom et al., 1956, p. 12). The taxonomy is divided into three domains: cognitive, affective, and psychomotor, however, the focus of this paper will be on the first domain. The six categories of the cognitive domain are (1) Knowledge, (2) Comprehension, (3) Application, (4) Analysis, (5) Synthesis, and (5) Evaluation. This taxonomy has been mostly used in the classification of learning goals and test items (Krathwohl, 2002).

Many studies critiqued and/or examined the validity of the taxonomy focusing, for instance, on the hierarchical nature of the taxonomy and the relationship between its levels (e.g., Coffman, 1956; Stanley & Bolton, 1957; Kropp & Stoker, 1966; Pring, 1971; Madaus, Woods, & Nuttall, 1973; Miller, Snowman, & O'Hara, 1979; Paul, 1993; Booker, 2007).

Anderson and Krathwohl (2001) revised Bloom's taxonomy and reported 12 changes, with the major changes being (1) the change of terminology and (2) the addition of a second dimension.

The first modification to the taxonomy was the renaming of some categories so they would be aligned with the vocabulary used by teachers. For instance, “Knowledge” was changed to “Remember”, “Comprehension” was changed to “Understand”, and “Synthesis” was renamed as “Create”. “Synthesis” also changed location with the “Evaluation” category. Another change in terminology was the adjustment of the categories’ names from a noun to a verb form. For instance, “Application”, “Analysis”, and “Evaluation” were changed to “Apply”, “Analyze”, and “Evaluate”, respectively. Table 2.1 gives the definitions of each category as reported in Krathwohl (2002).

Table 2.1

Bloom’s taxonomy definitions.

Category	Definition
Remember	Retrieving relevant knowledge from long-term memory
Understand	Determining the meaning of instructional messages, including oral, written, and graphic communication
Apply	Carrying out or using a procedure in a given situation
Analyze	Breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure or purpose
Evaluate	Making judgments based on criteria and standards
Create	Putting elements together to form a novel, coherent whole or make an original product

The second modification made to Bloom’s taxonomy was the inclusion of a second dimension named *Knowledge*. Originally, *Knowledge* was the first of the six categories under the Cognitive Process dimension, and it contained three subcategories: (1) knowledge of specifics, (2) knowledge of ways and means of dealing with specifics, and (3) knowledge of universals and abstractions in a field. The new *Knowledge* dimension of the revised taxonomy contain four categories, three of which are based on the three subcategories previously mentioned. The fourth additional category under the *Knowledge* dimension is called “metacognitive knowledge”. Krathwohl (2002), on page 214, defines the first category (factual knowledge) as “the basic

elements that students must know to be acquainted with a discipline or solve problems in it". The second category (conceptual knowledge) is defined as "the interrelationships among the basic elements within a larger structure that enable them to function together". Krathwohl defines the third category (procedural knowledge) as "how to do something; methods of inquiry, and criteria for using skills, algorithms, techniques, and methods", and the fourth category (metacognitive knowledge) as "knowledge of cognition in general as well as awareness and knowledge of one's own cognition".

The revised taxonomy also has a hierarchical structure with increasing complexity. However, the strict hierarchy is no longer present. Krathwohl (2002) stated that the changes made to the taxonomy allowed for overlap among categories, thus the requirement for a strict hierarchy was softened.

2.5.2 SOLO Taxonomy. The Structural Observation of Learning Outcomes (SOLO; Biggs & Collis, 1982) is a taxonomy developed based on students' learning and used to identify students' level of understanding. According to Biggs and Tang (2011) the SOLO taxonomy "classifies learning outcomes in terms of their structural quality, which makes it useful for defining levels of understanding, which in turn may be used for specifying such levels when writing learning outcomes" (page 81; Biggs & Tang, 2011).

SOLO is composed of five hierarchical levels, with each level being the basis for the next: (1) Prestructural, (2) Unistructural, (3) Multistructural, (4) Relational, and (5) Extended Abstract (Figure 2.2). Answers on the first level (Prestructural) present an incorrect response or no knowledge of the area. When an answer presents one relevant aspect, it can be categorized in the second level (Unistructural). Multistructural answers present many relevant aspects reported independently. Answers categorized as Relational contain many relevant aspects integrated in an

organized structure. Finally, when answers present generalizations to a new domain made from the organized structure, it can be categorized as Extended Abstract (Biggs & Collis, 1982; Biggs, 1989 and Boulton-Lewis, 1994). Biggs and Tang (2011) stated that the changes from the second to the third level are mostly *quantitative*, with more details being included in the responses. Moreover, the changes from the forth to the fifth level are mostly qualitative with the integration of concepts into a structural form (see Figure 2.2).

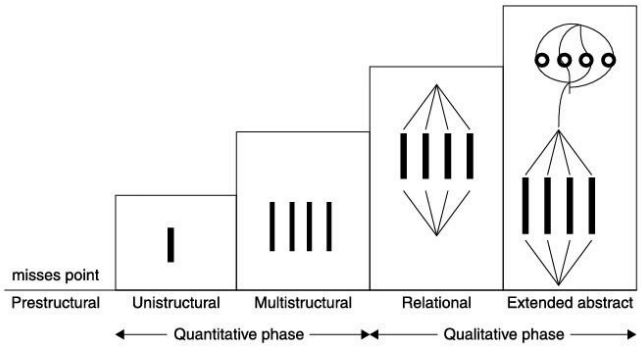


Figure 2.2. Modified from Biggs & Tang (2011), page 91.

Similar to Bloom's taxonomy, each level of SOLO has general verbs that can be used when writing learning outcomes for students. However, Biggs and Tang (2011) present different verbs depending on what type of knowledge will be addressed in the learning outcome. The authors report two categories of knowledge: declarative knowledge (or content knowledge) and functioning knowledge (knowledge that informs action by the learner). Table 2.2 lists some verbs for each of the knowledge categories.

Table 2.2

Verbs for knowledge categories. Modified from Biggs & Tang (2011), p. 124

	Declarative knowledge	Functioning knowledge
Unistructural	Memorize, identify, recite	Count, match, order.
Multistructural	Describe, classify.	Compute, illustrate.
Relational	Compare and contrast, explain, argue, analyse.	Apply, construct, translate, solve near problem, predict within same domain.

	Declarative knowledge	Functioning knowledge
Extended Abstract	Theorize, hypothesize, generalize.	Reflect and improve, invent, create, solve unseen problems, extrapolate to unknown domains.

Although Bloom's taxonomy has been widely used in education, there are very few references to this taxonomy in the statistics education literature (e.g., Garfield & Ben-Zvi, 2008). However, the SOLO categorization has been used more frequently in statistics education research (e.g., Watson and Moritz, 2000c, 2000d; Groth and Bergner, 2006; and Watson et al. 2003).

These taxonomies not only have been used to categorize learning outcomes but also to assess desired learning goals.

2.6 Assessments of Statistical Literacy and Reasoning

Assessments are used in research for many different purposes, such as to facilitate student learning, to provide feedback for students, to inform instructors regarding students' achievement, and to evaluate courses. National organizations such as the American Statistical Association (ASA, 2007), American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (AERA, APA, NCME, 1999) have outlined several suggestions for developing and improving instruments.

2.6.1 The Importance of Quality Assessments. The role of assessment in the field of statistics education is intrinsic not only in how students learn statistics but also in the teaching of statistics. The literature includes arguments against the use of final exam scores or course grades as indicators of statistical reasoning (e.g., Chance and Garfield, 2002; Konold, 1995). Despite all of this information available in the literature, Zieffler, Garfield, Alt, Dupuis, Holleque, and Chang (2008) indicate that there are still many studies using these measures.

Assessments can provide very important information related to students' learning but it is important to use quality instruments to capture this information. In a report published by the American Statistical Association (ASA, 2007) the authors suggested that every assessment should develop and report (1) information about the construct that is measured by the assessment, how the construct is aligned with the desired learning goals, and the limitations of the instrument; (2) information regarding the population of interest that the assessment will be administered to, the circumstances of administration or implementation of the assessment, and ways in which these are similar to or different from the setting in which published validity, reliability, and fairness evidence (if any) was obtained; and (3) evidence of validity, reliability, and fairness that is specific to the setting in which the assessment is administered, the particular population to which it is administered, the way it is scored, and the use to which the scores are put.

A clear definition of and differences between the learning outcomes can also help in the development of assessments (Garfield & Ben-Zvi, 2008). As mentioned previously, an important characteristic of an instrument is that it is aligned with the desired learning outcomes. One way of making this alignment is to create a test blueprint. A test blueprint lays out the content of the test and its relationship with the construct being measured by the test. Despite the challenges present in the assessment area in the field of statistics education, there has been development of instruments assessing the current learning goals for introductory statistics courses.

2.6.2 Existing Assessments. According to Garfield and Ben-Zvi (2008), changes are needed in how students are assessed due to the adoption of statistical literacy, reasoning, and thinking as new learning outcomes. Therefore, new instruments were developed to assess these learning outcomes. Currently, there are four assessments available in the field of statistics education

that measure statistical literacy and statistical reasoning (there is no published instrument available to assess statistical thinking yet).

Statistical Reasoning Assessment (SRA). The *Statistical Reasoning Assessment* was created from the need to evaluate a computer-based statistics curriculum developed as part of the ChancePlus Project (Konold, 1990 and Garfield, 1991) funded by the National Science Foundation (NSF Grant MDR-8954626). It was the first paper and pencil instrument developed and validated to assess some aspects of high school and college students' statistical reasoning (Garfield & Chance, 2000). Prior to SRA, most instruments and items used to assess students were focused on calculations and not on the students' ability to think and reason when facing statistics and probability problems. Therefore, SRA introduced a new era of assessments in the field of statistics education and until today it is the only instrument that provides not only measures of correct reasoning but also measures of incorrect reasoning.

Development. According to Garfield (1998b, 2003), the SRA is composed of 20 forced-choice items that cover specific types of reasoning and misconceptions. Liu (1998) explained that some items in the test already existed in the literature: two items were adapted from Konold (1989), three items from Kahneman & Tversky (1972), one item from Lecoutre (1992) and one item from Green (1983). The remaining 13 items were developed by the authors of the test and revised based on expert feedback. The forced-choice items presented in this instrument address statistics and probability problems. For each question, correct and incorrect alternatives are presented for the students to choose from (there can be one or more correct alternatives in each question).

The SRA does not assess all types of statistical reasoning. For example, formal and informal inferential reasoning are not assessed on the SRA. Instead, it covers the following types of correct reasoning: reasoning about data and representations of data, statistical measures, uncertainty, samples and association (Garfield, 1998b). The SRA instrument also assesses students'

misconceptions founded by many researchers (Kahneman, Slovic, & Tversky, 1982; Garfield & Ahlgren, 1988; Shaughnessy, 1992; Konold, 1989, 1995; Lecoutre, 1992) to be present in students' statistical and probabilistic reasoning. The misconceptions assessed by this instrument are related to averages, the outcome orientation, the idea that an effective sample should reflect as much as possible the population, the Law of small numbers, representativeness and equiprobability bias (Garfield 1998b, 2003).

This instrument provides eight scales for correct reasoning which when summed up yield a total correct reasoning score and eight scales for incorrect reasoning which when summed up yield a total score for incorrect reasoning.

Validity Evidence. Garfield (1998b, 2003) and Garfield and Chance (2000) presented information related to the validity evidence gathered to support the intended uses and inferences of the SRA test score. Content validity evidence was collected, using expert opinions and suggestions, and changes were made to the test items based on this information.

Response process validity evidence was also collected by analyzing individual responses from test pilots studies and constructed-response questions. Using all information gathered from experts, pilot studies, students' responses to constructed-response questions and different administrations of the test, the authors made many revisions, some exclusions or inclusions of items, until the final version of the SRA test was obtained.

Garfield (1998b, 2003), Garfield and Chance (2000), and Liu (1998) found that there was weak evidence of internal consistency of the test. Another estimate for the reliability (test-retest) of the SRA test scores was calculated based on the performance of 32 students enrolled in an assessment course at the University of Minnesota (Liu, 1998). The SRA instrument was used as a pre-test and after a week as a post-test yielding a test-retest estimate of reliability of 0.70 and 0.75 for correct and incorrect reasoning, respectively.

The authors of the test also collected validity evidence regarding relation to other variables (referred by the author as criterion-related validity). Based on the evidence gathered from correlating the final test scores (correct and incorrect reasoning) from SRA with some course outcomes such as final scores at the end of a first course of statistics, Garfield (2003) suggests that there is no relationship between outcomes in an introductory statistics course and students' correct and incorrect reasoning.

Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS). The *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS, delMas et al., 2007) was designed to assess students' statistical reasoning after taking an introductory statistics course (delMas et al., 2007). This instrument is composed of 40 forced-choice items that assess students' conceptual understanding of data collection and design, descriptive statistics, graphical representations, boxplots, normal distribution, bivariate data, probability, sampling variability, confidence intervals, and tests of significance (delMas et al., 2007). Although the CAOS items were written to assess students' reasoning on content typically covered in an introductory course, primarily the items focused on assessing variability (Garfield, delMas, & Chance, n.d.).

Development. The CAOS test was developed over the course of three years (see delMas et al., 2007 for more detail). Many of the items for CAOS were written by the tests developers. Others were obtained from an item database created as part of the Assessment Resource Tool for Improving Statistical Thinking (ARTIST) project funded by the National Science Foundation (NSF DUE-0206571). The first version of the CAOS test—CAOS 1—was revised after it was piloted in August 2004. At this time, CAOS 1 consisted of 34 forced-choice items, but after the revision 37 forced-choice items were included—CAOS 2.

In the beginning of 2005, CAOS 2 was used as an online and in-class instrument for approximately 900 students (from secondary and post-secondary level) and the results from the

analysis of the students' responses were used to make more improvements in the test, leading to the next test version—CAOS 3. This third version was reviewed by 30 statistics instructors, and revisions were made based on the instructors' feedback leading to the fourth and final version of the CAOS test—CAOS 4—with 40 forced-choice items.

Validity evidence. Validity evidence was gathered to support the intended inferences and uses of the CAOS scores. Each of the CAOS items were reviewed by experts (ARTIST advisory board). These experts reviewed the items for factual context and data, as well as, for adherence to item writing guidelines. Eighteen experts agreed that they would expect their students to attain the outcomes measured by the instrument and that the CAOS test "measures basic outcomes in statistical literacy and reasoning that are appropriate for a first course in statistics" (delMas et al., 2007). The ARTIST advisory board also provided feedback regarding important concepts that were not being measured by the items. New items were developed and other items were revised as necessary based on the feedback received from the experts. In addition to the feedback provided, experts' validity ratings of the CAOS items were also collected. These sources provided content validity evidence during each phase of the test development.

Response process validity evidence was also collected for the CAOS test. This evidence came from analyzing students' responses collected during several pilot studies. The information gathered from the pilot studies was used, for example, to remove response items which were not selected by students. These studies also contributed to identify items that were not functioning well or did not contribute to the scale. Such items were eliminated, modified or replaced by new items (R. delMas, personal communication, October 7, 2012).

Additional validity evidence regarding score precision of the CAOS test was also obtained. In a study to compare statistical reasoning at the beginning and end of an introductory statistics course, delMas et al. (2007) estimated a reliability coefficient (coefficient alpha) of 0.82 using a

sample of 1,470 students. The students included in the sample were from introductory statistics courses with many different mathematics pre-requisites. Zieffler (2006) also reported an estimate for the coefficient alpha of 0.81 using a sample of 110 undergraduate students.

Goals and Outcomes Associated with Learning Statistics (GOALS). The *Goals and Outcomes Associated with Learning Statistics* (GOALS) instrument was developed to assess important statistical reasoning outcomes in a first course of statistics. GOALS is currently on its fourth version (GOALS-4) which is composed of 20 items that address the topics of study design, bivariate relationships, variability, sampling and sampling variability, interpreting confidence intervals and p-values, statistical inference, and modeling and simulation.

Development. The GOALS instrument was developed as an updated version of the *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS, delMas et al., 2007) test. Similar to CAOS, the GOALS instrument assesses student statistical reasoning in a first course of statistics. One of the primary reasons the GOALS instrument was created was to evaluate the effectiveness of the CATALST curriculum in developing students' conceptual ideas and statistical reasoning. In its first iteration, the authors of GOALS analyzed students' responses to CAOS using both distractor analysis and item response theory. The results of these analyses were used to identify CAOS items that were not performing well and delete them from the instrument. Additionally, other items were modified based on research done by Ziegler (2012), who explored how the stem length of forced-choice items affected students' responses.

At this point, the GOALS items were examined for alignment with current learning goals for introductory statistics courses. This suggested a need for items that addressed topics such as a simulation and randomization approach to inference. It also suggested that items that addressed topics such as the purpose of random assignment and the interpretation of statistically significant results and *p*-values were needed. As a result, 14 items were added to the GOALS instrument. Of

these 14 items, seven were adapted from the NSF-funded *Concepts of Statistical Inference* project (CCLI DUE-0633349), three were adapted from the NSF-funded *Creating a Teaching and Learning Infrastructure for Introductory Statistics Re-design* project (NSF DUE-0737126), and four were developed by University of Minnesota faculty. This resulted in the first version of the GOALS instrument (GOALS-1), comprised of 28 items—three constructed-response items and 25 forced-choice items.

The GOALS instrument was modified based on the feedback received from the reviewers and discussion with another expert in the field of statistics education, leading to its second version-GOALS-2. One of the changes was the development of two forms of the instrument. This decision, motivated by feedback from content experts, was made because the assessment of students' reasoning about inference seemed tied to whether students learned inference using a classical or simulation/randomization-based approach. Subsequently, one form was developed for students enrolled in statistics courses with classical content and another form was developed for students enrolled in simulation- and randomization-based courses (e.g., CATALST). Each of the two forms of the GOALS instrument was comprised of 27 forced-choice items. Apart from the four items with content related to the use of simulation to carry out statistical inference, the remaining 23 items on the two forms are identical.

In a field-test during fall 2011 and spring 2012, the simulation/randomization form of the GOALS instrument was administered to 289 undergraduate students from six universities in the United States. These students were enrolled in a statistics course that was using the CATALST curriculum. The CATALST curriculum was designed as part of the NSF-funded CATALST Project (NSF DUE-0814433), which developed materials, lesson plans, and assessments based on a simulation/randomization-based approach to statistical inference. The classical form of the instrument was also administered, during the same time, to 569 students in consensus introductory

statistics courses at six universities in the United States. Psychometric analysis of the field-test data, using classical test theory and item response theory, led to the deletion of some items, and modification of other items that were not discriminating well among students with high and low ability. In addition, after meetings among the developers of GOALS, it was decided that the instrument would be composed of only one form. As a result, the third version of GOALS (GOALS-3) was obtained.

Another field-test was performed during spring 2014, when the GOALS-3 instrument was administered to 932 students in introductory statistics courses at 21 institutions (colleges and high schools) in the United States. Additional psychometric analysis was performed which led to the deletion of one item and the addition of a new item. The GOALS instrument is now on its fourth version (GOALS-4) and is now composed of 20 items.

Validity evidence. In order to provide validity evidence about the content assessed by the GOALS-2 instrument, statistics instructors doing research in the field of statistics education were identified and invited to provide feedback about the GOALS items. Seven instructors accepted this invitation. Reviewers were requested to rate the extent to which they agreed that each GOALS item measured an important learning outcome for any student who had completed a college-level, non-calculus based, introductory statistics course. Reviewers were also asked to list any learning outcome that was not assessed by the GOALS instrument but at the same time was an important learning outcome for them. Finally, the reviewers were asked to identify and offer modifications for any items that they felt needed improvement.

After analysis of field data in 2013, surveys with statistics instructors were done to provide additional validity evidence regarding the relationship between the items in GOALS-3 and the unidimensional construct being measured by the instrument.

Evidence based on the internal structure of the GOALS test was also obtained. Confirmatory factory analysis of field-test data showed that the GOALS-2 scores were measuring a unidimensional construct. Additional evidence of validity for GOALS-2 and GOALS-3 was collected through psychometric analysis of field data under the item response theory framework.

The GOALS instrument is still under development and its fourth version (GOALS-4) is currently being administered to introductory statistics courses in the United States. Analysis of data from the current field-testing will lead to additional validity evidence to support the intended inferences and uses of the GOALS-4 test scores.

Basic Literacy in Statistics (BLIS). The *Basic Literacy in Statistics* (BLIS; Ziegler, 2014) assessment was created to measure students' ability to read, understand, and communicate statistical information. This instrument was developed due to the need for an assessment that measures students' statistical literacy and that is aligned with curriculums that use randomization and simulation methods. The assessment was designed to assess statistical literacy of students enrolled in introductory statistics courses at the postsecondary level that include some coverage of simulation-based methods in the curriculum. The topics addressed by BLIS are (1) data production, (2) graphs, (3) descriptive statistics, (4) empirical sampling distributions, (5) confidence intervals, (6) randomization distributions, (7) hypothesis tests, (8) scope of conclusions, and (9) regression and correlation.

Development. Ziegler (2013) reported that the definition of statistical literacy used to develop this instrument was "the ability to read, understand, and communicate statistical information". The BLIS instrument is composed of 37 items assessing the 9 topics described above.

According to Ziegler (2013), items from BLIS were a mixture of items from CAOS, items from an earlier version of GOALS, and items from the ARTIST Topic Scale tests and ARTIST item database.

Ziegler (2013) reported that the development of the BLIS instrument started with a test blueprint that was modified after being reviewed by statistics instructors. One of the suggestions was the addition of a topic about regression and correlation. The new version of the blueprint contained 37 learning outcomes and it led to the preliminary version of the BLIS assessment, which contained multiple-choice and constructed-response items. This version was also reviewed by the same statistics instructors and based on their feedback, some items were modified to better align with their proposed learning outcome. These changes resulted in the first version of the instrument (BLIS-1).

In the next phase of the development, think aloud interviews were performed and analyzed. Based on the results from the interviews, changes were made to 15 items. BLIS-2 was the result of these changes. A pilot study was also carried out and was used for conversion from constructed-response items to multiple-choice items. The result of the analysis of the pilot-data also led to the update of one of the items, which was considered very difficult. These changes led to the third and final version of the assessment (BLIS-3).

Validity evidence. Expert review provided content validity evidence for the intended uses of the BLIS's scores. Six statistics educators reviewed the preliminary test blueprint and were asked to rate how important each learning outcome was in determining how statistically literate a student was. Reviewers were also asked to describe other topics and learning goals that were not included in the blueprint but were related to Ziegler's definition of statistical literacy. The same statistics educators reviewed the preliminary version of the instrument.

Response process validity evidence was collected as part of the development process. Cognitive interviews with six students, from four different introductory statistics courses, were conducted to understand students' thinking when answering the items from the test. A pilot study was also carried out with 76 students in three introductory statistics course. From the analysis of

students' responses, changes were made in the instrument, such as conversion from constructed-response items to selected-response items. Students' answers from the pilot data were used to create plausible distractors for the selected-response items.

Evidence based on the internal structure of the test was also obtained. BLIS-3 was used in field-testing with 940 students from 34 introductory statistics courses. Thirty-two statistics courses from the United States, one statistics course from Canada, and one statistics course from Spain. Confirmatory factory analysis of field-test data was performed and revealed that the scores were measuring a unidimensional construct. According to Ziegler (2013) data analyses based on classical test theory and item response theory were used to provide additional validity evidence.

Finally, this same data from the field test was used to gather validity evidence regarding score precision. Ziegler reported that the coefficient alpha was .83, indicating adequate score precision.

Even though there are current assessments of statistical literacy and reasoning, it is important to consider that these instruments could have different psychometric characteristics depending on the content and method of instruction of introductory statistics courses.

2.7 Discussion

A review of the literature was presented about defining, assessing and distinguishing the constructs of statistical literacy, reasoning, and thinking. Information about taxonomies for classification of educational goals was also reported. In the following section, a critique of the literature is presented followed by the problem statement and research questions.

2.7.1 Critique of the Literature Reviewed. The literature reveals challenges in defining and assessing the learning goals of statistical literacy, reasoning and thinking as well as in

understanding their relationship. There are open questions regarding the possible hierarchy formed by these constructs.

There is a lack of agreement among the current definitions of statistical literacy, statistical reasoning, and statistical thinking. Since these constructs are important learning goals for students, there is a need to have a more consistent definition of these terms. A unified definition of statistical reasoning, literacy, and thinking would benefit statistics education, by providing agreement regarding these terms. Currently, most of those who define a concept do it through their own lenses, based on their own knowledge, with little empirical evidence to support their definitions.

A second concern is related to the instruments assessing these learning goals. Currently, there is no assessment of statistical thinking. In addition, each of the instruments available measure the constructs of statistical literacy and statistical reasoning separately. However, it seems that there may be some dependency between these constructs. Therefore, there is a need to develop an instrument assessing both statistical literacy and reasoning concurrently. It is also important to notice that the SRA, CAOS and GOALS assessments did not have a clear working definition of statistical reasoning which was the construct being measured by these tests.

There appears to be a lack of clarity and evidence regarding the relationship among statistical literacy, reasoning, and thinking; how much they overlap (if at all). For instance, Rumsey (2002) defined "statistical citizenship" (which is considered a part of statistical literacy) and mentioned that students take actions that may require statistical reasoning, such as the judgment and evaluation of statistical information. However, Rumsey (2002) also stated that "statistical competence" (which is also considered as a part of statistical literacy) is a requirement for statistical reasoning. Therefore, statistical reasoning seems to partially overlap with statistical literacy. Another example of this overlap in the literature is found in Budgett and Pfannkuch's (2007) definition of adults' statistical literacy which contains a statistical reasoning component. Therefore,

it is not clear if statistical literacy partially overlaps with statistical reasoning or if statistical literacy is a pre-requisite for statistical reasoning. Although the relationship between these two concepts was examined by delMas (2002), more research is needed to investigate which model better represents the connections between statistical literacy and reasoning.

An overlap may also involve statistical thinking (Garfield & Ben-Zvi 2007, 2008), which may contain a statistical reasoning component. In addition, the statistical thinking framework obtained by Wild and Pfanchuch (1999) was based on interviews which aimed to uncover the statisticians' and students' *statistical reasoning* processes. Therefore, statistical reasoning is being used to understand statistical thinking, which can be considered additional evidence for the existence of overlap.

Another issue identified in the literature concerns the hierarchy between statistical literacy, reasoning and thinking and a possible alignment between these three learning goals and Bloom's and SOLO taxonomies. For instance, statistical literacy, reasoning, and thinking were compared to some of the categories in Bloom's taxonomy (Garfield & Ben-Zvi, 2008). Statistical literacy aligned with the *Knowledge* category; statistical reasoning aligned with the *Comprehension* category and partially aligned with the *Application* and *Analysis* categories; and, finally, statistical thinking corresponded to many aspects of the *Application*, *Analysis*, *Synthesis*, and *Evaluation* categories. Despite this comparison, no empirical information was reported regarding this possible alignment with Bloom's taxonomy. In addition, the hierarchy between statistical literacy, reasoning, and thinking is not supported by any empirical evidence in the literature. Some verbs used to create learning goals under the SOLO framework partially match the verbs used by delMas (2002) to develop assessment items to measure statistical literacy, reasoning, and thinking. In spite of these similarities, no work has been done regarding a possible alignment between statistical literacy, reasoning, and thinking and the SOLO taxonomy.

2.8 Problem Statement and Research Questions

Many unanswered questions exist related to defining, assessing and understanding the structure of statistical literacy, reasoning and thinking; therefore, research on the connections between these constructs would benefit the field of statistics education. However, due to the lack of assessments available to measure students' statistical thinking, this study focused only on statistical literacy and reasoning.

As mentioned previously, statistical literacy and statistical reasoning have been defined with little empirical evidence and questions remain regarding the possible hierarchy and overlap between these two learning goals. To address these problems, the following research questions were investigated in this study: (1) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony? (2) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction? (3) What measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning?

A unique aspect of this study is that it needs to examine the selection and fit of a model according to a statistical approach and a measurement approach. In a statistical approach, the model selection is judge solely on statistical measures of fit (e.g. S-X^2, RMSEA, AIC, and BIC) and parsimony (i.e., selecting the model with the fewest parameters that best fits the data). However, a measurement approach focuses on the usefulness of the models in distinguishing between the constructs and in measuring the constructs reliably. The third research question builds on the results from the previous two research questions by selecting a model that has evidence of good statistical fit and evidence of practical utility in the use of subscores.

Chapter 3

Methods

In order to answer the research questions posed in the previous chapter, a preliminary version of a new assessment was developed that measures both statistical literacy and statistical reasoning. Data collected using this assessment was used to answer the research questions.

This chapter starts with a brief overview of the study followed by working definitions of statistical literacy and statistical reasoning. Next, information is provided regarding the development of the new assessment. The assessment development process includes elaboration of a test blueprint, expert revision, think-aloud interviews with students followed by a pilot study, and finally the field test. The chapter ends with the description of the field test's data analysis and a summary of the chapter.

3.1 Study Overview

The purpose of this study was to evaluate the relationship between statistical literacy and statistical reasoning. Specifically, the research questions investigated were the following: (1) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony? (2) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction? (3) What measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning?

In order to answer these research questions, there was first a need to measure the learning goals of statistical literacy and statistical reasoning. Thus a new instrument measuring these learning goals concurrently was developed. All steps of the instrument development were reviewed

by the author's co-advisers at the University of Minnesota, Dr. Joan Garfield and Dr. Andrew Zieffler.

The development process of the new assessment began with establishing the working definitions of statistical reasoning and statistical literacy used throughout the study. Once the definitions were set, a test blueprint was created based on the content addressed by two existing assessments: the BLIS assessment (Ziegler, 2014) which measured students' statistical literacy and the GOALS assessment (Sabbag & Zieffler, 2015) which measured students' statistical reasoning. The first draft of the instrument was mostly composed by items from these two assessments. However, additional items were also used from other existing instruments and the ARTIST item bank.

Identification of behaviors needed to answer each item correctly was also part of the development process. Based on these behaviors and the working definitions of statistical literacy and statistical reasoning, items from the first draft of the instrument were categorized as statistical literacy or statistical reasoning items. The categorization of items into these two categories was also done by experts in the field of statistics education (expert review). In addition, experts were also asked to critique the items. Feedback from the experts was used to create the second draft of the instrument. In addition to expert feedback, think-aloud interviews were also conducted with students to better understand their response process. Information based on these interviews was used to make changes to the items which led to the third draft of the instrument.

A pilot test was conducted using the third draft of the instrument and the pilot results were used to improve items which then became the first version of the assessment. This version was then field tested and data analysis was conducted to answer the research questions posed for this study. The timeline for the study is shown in Table 3.1.

Table 3.1

Timeline of the study

Procedures	Timeline
Development of the preliminary version of the assessment	September – October 18, 2015
Invite reviewers to participate in study	September 28, 2015
Expert review of preliminary version	October19 – November 9, 2015
Think-aloud interviews	November 15 – November 25, 2015
Pilot study	December 1 – December 7, 2015
Field test	April, 2016
Analysis of data	May, 2016

3.2 Working Definitions of Statistical Literacy and Statistical Reasoning

For the purpose of this study, it was necessary to adopt definitions of statistical literacy and statistical reasoning. The main purpose of the definitions was to give clear guidance to differentiate statistical literacy items from statistical reasoning items. Therefore, it was decided to make the definitions of literacy and reasoning focused on items.

The definition of statistical literacy that was used in this study was based on the definition adopted by Ziegler (2014): *Statistical literacy items assess students' ability to recall a definition, describe or interpret basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will not require that students make connections between them (recall information will be sufficient).*

Statistical reasoning was defined based on the definitions from Garfield and Ben-Zvi (2008) and delMas (2002, 2004): *Statistical reasoning items assess students' ability to make connections among statistical concepts, create mental representations of statistical problems, and explain relationships between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, statistical reasoning items require higher order thinking and higher cognitive load than statistical literacy items.*

3.3 Instrument Development

The instrument that was used in this study was initially composed of questions from the *Goals and Outcomes Associated with Learning Statistics* (GOALS; Sabbag & Zieffler, 2015) assessment and questions from the *Basic Literacy in Statistics* (BLIS; Ziegler, 2014) assessment. This new instrument will be referred to as REALI (REasoning and Literacy Instrument) for the remainder of this paper.

As mentioned earlier, there are currently three assessments that were designed to measure students' statistical reasoning: *Statistical Reasoning Assessment* (SRA-Garfield, 1991), *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS-delMas et al., 2007), and GOALS (Sabbag & Zieffler, 2015). Because GOALS is the instrument which has the most validity evidence to support the intended uses of the scores, the items from GOALS were used in the REALI assessment. The internal consistency estimate of reliability (coefficient alpha) for the GOALS instrument is 0.67; however, the instrument is currently being updated to improve its psychometric characteristics. The GOALS instrument is composed of 20 items that address topics such as study design, sampling variability, interpreting *p*-values, and statistical inference.

The BLIS assessment (Ziegler, 2014) was created to measure students' ability to read, understand, and communicate statistical information (statistical literacy definition) and it is composed of 37 items. This is the only published assessment measuring students' statistical literacy. Extensive validity evidence to support the intended inferences and uses of the BLIS' scores was reported by Ziegler (2014). Specifically, she reported that the internal consistency estimate (coefficient alpha) for reliability was 0.83.

The REALI assessment was initially composed of items from GOALS and from BLIS and the aim was to have 20 items measuring statistical literacy and 20 items measuring statistical reasoning, resulting in an instrument with 40 items. The rationale for having each part of the test

composed of 20 items is based on advice from Sinharay (2010). Sinharay reported that for subscores to have added value over the total scores they need to be composed of at least 20 items.

3.3.1 Blueprint. To obtain the blueprint of the assessment, the primary concepts addressed by each item from BIS and from GOALS were identified. Table 3.2 shows all concepts addressed by both instruments and the number of items in each instrument measuring each of the statistical concepts.

Table 3.2

Number of BLIS and GOALS items for each statistical concept

Concept	# GOALS items	# BLIS items	Total
t-test	1	0	1
p-value	5	3	8
Conclusion based on p-value/significance	3	3	6
Design/conclusion	3	4	8
Correlation	1	1	3
Graph	3	5	12
Outlier	2	3	5
Mean/median	3	4	10
Standard deviation/variability	7	9	17
Hypothesis testing	6	11	16
Null hypothesis	2	6	8
Sample size	5	3	9
Confidence interval	3	4	7
Simulation/randomization	2	3	5
Randomness	0	1	1
Probability	0	2	2
Types of variables	0	1	1
Characteristics of study	0	4	4
Sampling	1	2	3
Characteristics /describe distribution	3	4	3
Sampling/randomization distribution	3	7	8
Type errors	0	2	2
Regression	0	1	1

Items that assessed similar content were grouped into eight areas of learning: (1) representations of data, (2) measures of center, (3) measures of variability, (4) study design, (5) confidence intervals, (6) hypothesis testing & *p*-values, (7) probability, and (8) bivariate data. The first three columns of Table 3.3 show the areas of learning and the number of GOALS and BLIS items in each area. See Appendix A for detailed information regarding which items were assigned to each of the areas of learning. The learning goals for each of the items are also shown in this appendix.

Table 3.3

Number of BLIS and GOALS items and percentage for each area of learning.

				Target		
Area of learning	**BLIS items**	**GOALS items**	**Total**	**Statistical Literacy**	**Statistical Reasoning**	**Total**
Representations of data	4	1	5 (9%)	2	2	4
Measures of center	2	1	3 (5%)	2	2	4
Measures of variability	3	2	5 (9%)	2	2	4
Study design	10	3	13 (23%)	3	3	6
Confidence intervals	4	3	7 (12%)	2	2	4
Hypothesis testing & *p*-values	10	9	19 (33%)	5	5	10
Probability	2	0	2 (4%)	2	2	4
Bivariate Data	2	1	3 (5%)	2	2	4
Total	37	20	57 (100%)	20	20	40

As shown in Table 3.3, each area of learning contained different numbers of BLIS and GOALS items. However, to have a balanced instrument, it was the aim of the study to have the same number of statistical literacy and reasoning items within the areas of learning. The rationale used to decide the desired number of items of each area of learning was the following. The fourth column in Table 3.3 shows the total number of items for each area of learning across instruments. It can be seen that the *hypothesis testing & p-values* area of learning had the highest proportion of items (33%) followed by *study design* (23%). Therefore, it was decided to assign a higher number of items to these two areas of learning and leave all remaining six areas with the same number of

items. To accomplish this, it was established that the *hypothesis testing & p-values* area of learning would have a total of 10 items (five statistical literacy and five statistical reasoning items), *study design* would have a total of six items (three statistical literacy and three statistical reasoning items), and all six remaining areas of learning would have a total of four items each (two statistical literacy and two statistical reasoning items). The last columns of Table 3.3 show the desired number of items for each area of learning.

The next step to obtain the blueprint of the REALI instrument was the categorization of items into statistical literacy or statistical reasoning. This process was necessary because the items from BLIS (technically statistical literacy items) and from GOALS (technically statistical reasoning items) were not developed under the working definitions used in this paper. Therefore, it was necessary to check the alignment of each item with the working definitions of statistical literacy and statistical reasoning.

Before the classification of the items, it was necessary to identify the behaviors needed to answer each item correctly. To identify the behaviors, the author of this paper read each item and listed what were the abilities and understanding needed by a student to answer each item correctly. An example of an item and its behavior is given in Figure 3.1. Appendix B shows the behaviors for each item from BLIS (Appendix B1) and from GOALS (Appendix B2).

Based on the behaviors identified for each item and on the working definitions, each item was categorized as a statistical literacy item or a statistical reasoning item. At the end of the classification, there were two items (Items 3 and 35) from BLIS that were classified as statistical reasoning items and four items (Items 6, 16, 17, and 18) from GOALS that were classified under statistical literacy items. All these items were not included in the REALI assessment. Some items were not clearly categorized as a reasoning or as a literacy item. These items were further investigated and discussions with two faculty members at the University of Minnesota were done

to reach an agreement regarding the items' classification. These items were also included in the think-aloud interview with students so that more information could be gathered that could help in the categorization of these items.

ITEM:
The Pew Research Center surveyed 2,076 American adult cell phone users chosen at random in 2013. The sample percent of adult cell phone users who access the internet on their phones was 60%. The 95% confidence interval was 58% to 62%. What is this interval attempting to estimate?

a. The average number of American adult cell phone users who access the internet on their phones in 2013.
b. The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
c. The percent of all American adult cell phone users who access the internet on their phones in 2013.
d. For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.

BEHAVIORS:
To answer the item above correctly, students need to
1) Understand what a confidence interval represents.
2) Recognize which parameter is being estimated.
3) Recognize the population of interest.
4) Understand what the level of confidence represents

Figure 3.1. Statistical literacy item and behaviors.

Several additional items were deleted because they were assessing content specific to randomization-based curricula (e.g., simulation and randomization distributions). For instance, six items (Items 18, 19, 23, 24, 25, and 26) from BLIS and four items (Items 15, 16, 17, and 18) from GOALS were deleted due to the content they were assessing. One item (Item 20) from GOALS was modified to no longer address a simulation topic. The reason for not using items with very specific content is because this instrument was designed to serve all types of introductory statistics courses.

In some areas of learning the number of items exceeded the required number established by the author. For these areas, some items were deleted. The decision for deleting items was based

on (1) psychometric information (item discrimination and item difficulty) available from previous analysis and (2) the alignment of the content addressed by the item and the content of the other items included in the instrument. For instance, some items measured very similar content so items with the worst item characteristics were deleted. Table 3.4 shows the number of working items for each area of learning, after the process of categorizing and deleting items.

As can be seen in Table 3.4, after the process of categorization and verification of items, some areas of learning did not have the required number of items. The next section reports the process of adding new items to these areas of learning.

Table 3.4

Statistical literacy and reasoning items for each area of learning.

Area of learning	Statistical Literacy items	Statistical Reasoning items	Target	
			Statistical Literacy	Statistical reasoning
Representation of data	2	1	2	2
Measures of center	2	1	2	2
Measures of variability	2	2	2	2
Study design	3	3	3	3
Confidence intervals	2	2	2	2
Hypothesis testing & *p*-values	5	2	5	5
Probability	2	0	2	2
Bivariate Data	2	1	2	2
Total	20	12	20	20

3.3.2 Additional items. New items were added to six areas of learning (representations of data, measures of center, study design, hypothesis testing & *p*-value, probability, and bivariate data) so that each area would contain the same number of statistical literacy and statistical reasoning items. These new items originated from existing instruments (e.g., CAOS, AIRS), from the ARTIST Topic Scale item bank, or from materials used in introductory statistics courses at the University of Minnesota. Table 3.5 shows the number of working items for each area of learning,

after new items were added. This pool of items contained 27 statistical literacy items and 25 statistical reasoning items, for a total of 52 items. This was the first draft of the REALI assessment (see Appendix C).

It can be seen that with the addition of new items, this first draft of the instrument had more items than needed. This was allowed because there was no previous psychometric information regarding these new items. Therefore, there was a need to first verify the characteristics of such items through the pilot test and then make a decision regarding which items should be retained and which should be deleted. This first draft of the assessment was used in the expert review which will be explained in the next section.

Table 3.5

Number of statistical literacy and statistical reasoning items for the first draft of the instrument

Area of learning	**Statistical Literacy items**	**Statistical Reasoning items**
Representation of data	2	3
Measures of center	3	2
Measures of variability	2	2
Study design	3	5
Confidence intervals	8	6
Hypothesis testing & *p*-values	2	2
Probability	3	3
Bivariate Data	4	2
Total	27	25

3.3.3 Expert Review. To provide additional support regarding the categorization of items into statistical literacy or statistical reasoning, an expert review was conducted in two phases.

Phase 1. The reviewers for the initial phase were four faculty from the University of Minnesota (Robert delMas, Elizabeth Fry, and Laura Le) and one faculty (John Holcomb) from Cleveland State University. These scholars were chosen because they had done work or were interested in assessment development and statistical learning goals. The intent of this first phase

was to observe how the reviewers would categorize the items and to verify if there would be any difficulties in the categorization process. Therefore, the data from this first part of the review was not used in the study.

The four reviewers were invited to participate in the study by email (see Appendix D1) and all of them agreed to participate. The author and the reviewers exchanged emails and agreed on a date and time to meet in person. In the personal meeting, the reviewers were given the working definitions of statistical literacy and statistical reasoning used in this study. However, because the terms "statistical literacy" and "statistical reasoning" are widely used in the field of statistics education and people have different opinions about them, the author decided to not include the names "statistical literacy" and "statistical reasoning" in the definitions to decrease bias. Instead the working definitions used the terms "GROUP 1" and "GROUP 2":

- Items in GROUP 1 assess students' ability to recall, describe or interpret basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will not require that students make connections between them (recall information will be sufficient).
- Items in GROUP 2 assess students' ability to make connections among statistical concepts, create mental representations of statistical problems, or explain relationships between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

In addition to the working definitions, the reviewers were also provided with 52 pieces of paper each containing one of the 52 items. All items were grouped according to their area of

learning, but the order of the items within each area of learning was randomized. In addition, the order of the eight areas of learning was also randomly chosen for each of the four reviewers. The reviewers were then verbally instructed by the author to categorize each of the items into one of the two groups. During the categorization process, the reviewers were asked to verbalize their thinking process.

Based on the performance of the reviewers, the author made slight changes to the working definitions, such as bolding, underlining, and italicizing some of the letters to call attention to important parts of the definitions. In addition, the author also decided to include the words "a definition" after the verb *recall* to be more explicit about what the verb was referring to. The final working definitions are shown below and these were the definitions used in the second phase of the expert review.

- Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).
- Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

Phase 2. The second phase of the expert review was conducted with four experts in the field of statistics education (Maxine Pfannkuch, Dani Ben-Zvi, Jane Watson, and Rob Gould) who had done work or were interested in statistical literacy, statistical reasoning, or statistical thinking.

An email was sent to all four reviewers asking if they were willing to participate in the study by doing the categorization of items (see Appendix D2). All agreed to participate and were sent the review form by email (see Appendix D3). The form contained (1) the working definitions of statistical literacy and statistical reasoning (again without using the words "statistical literacy" and "statistical reasoning") and (2) the preliminary version of the REALI assessment with 52 items. The reviewers were also asked to critique and provide feedback about the items.

Changes were made to some of the items based on the expert feedback and this led to the second draft of the REALI instrument. All changes were related to improving the clarity of the items or adjusting the length of the alternatives to follow item writing guidelines. These changes were approved and supervised by the author's co-advisers Dr. Joan Garfield and Dr. Andrew Zieffler. This same version of the REALI instrument was used in the think-aloud interviews.

3.3.4 Think-aloud Interviews. To provide evidence for how students were responding to the items (response process validity evidence), think-aloud interviews were conducted with students from the University of Minnesota. The target population of students for the think-aloud interviews was students who had recently taken an introductory statistics course (undergraduate and graduate level) or were currently taking a second course of statistics (undergraduate and graduate level).

To gather the participants, three University of Minnesota statistics instructors from the Education Psychology department were asked to email their students asking for volunteers to participate in the think-aloud interviews (see Appendix E1). One instructor emailed students who

had taken an undergraduate introductory statistics course in the previous semester. The second instructor emailed students who had taken a graduate level introductory statistics course in the previous semester. And the third instructor emailed his current students who were enrolled in a graduate level second course of statistics. It was the intent of the author to have one student from each of these three statistics courses. However, only students from the last two courses responded to the invitation emails.

In addition to inviting students taking statistics courses from the Educational Psychology department, the author also emailed instructors from the Statistics department at the University of Minnesota. An email was sent to two instructors who were currently teaching a second course of statistics for statistics majors, asking if the author could visit their class and ask for volunteers (see Appendix E2). Only one instructor for the Statistics department responded to the email and authorized the author to visit the class. It was the intent of the author to have at least one undergraduate student from the Statistics department.

The script for the in-class visit can be seen in Appendix E3. To encourage participation, the invitation email (and the in-class script) stated that a gift card would be given to each volunteer who participated in the think-aloud interview.

The first three students from the Educational Psychology courses who responded to the email invitation were selected to participate in the interview. In addition, only one student from the Statistics department volunteered to participate in the study. Therefore, a total of four students agreed to do the think-aloud interviews (one undergraduate and three beginning master's students). The author and the students exchanged emails to find a date and time to meet and do the interview. The interview was conducted in the author's office at the University of Minnesota.

In the beginning of the think-aloud interview, students were given the consent form (see Appendix E4) and asked to read and sign it. The author then read the interview protocol (see

Appendix E5) to the students which encouraged them to verbalize their thinking while answering the questions from the REALI assessment. Next, students were given the items from the REALI assessment and asked to read each item out loud and say what they were thinking while answering the items. The audio of these think-aloud interviews was recorded.

Since questions from BLIS and from GOALS had already been through a process of validation, the think-aloud interview only used new items, modified items, or items that could not be categorized with certainty as a statistical literacy or a statistical reasoning item. A total of 26 items were used in the think-aloud interviews (see Appendix D3).

Based on the students' responses, the researcher identified items that were not clear to the students. For instance, some items were not interpreted by the students in the way that the items were designed to be interpreted. In addition, students would sometimes read the same item many times or verbalize if an item was confusing or not clear. Modifications were made to the items based on students' responses and these modifications were approved and supervised by Dr. Joan Garfield and Dr. Andrew Zieffler (author's co-advisers). These changes led to the third draft of the REALI instrument which was used in the pilot study (see Appendix C2).

3.3.5 Pilot Test. The next step in the development process was a pilot study using the third draft of the REALI assessment. Five instructors (author included) from two introductory statistics courses at the University of Minnesota (EPSY 3264 and EPSY 5261) were contacted by email to administer the assessment to their students. In order to have students from other institutions, the researcher contacted by email instructors who usually attend a monthly meeting of statistics educators called Stat Chat (at Macalester College in Saint Paul, MN). One instructor from Stat Chat showed interest in participating in the pilot study.

A total of six instructors from the University of Minnesota and Augsburg College agreed to participate in the pilot study. An email was sent to each instructor giving information related to the administration of the instrument (see Appendix F1). To increase student's participation, all instructors administered the REALI assessment as an extra credit opportunity. The administration of the instrument was done online through Qualtrics. The initial page of the assessment contained a consent form asking if students were willing to participate in the study. Information contained in the consent form clearly explained to the students that they would still receive extra credit even if they did not want to give consent (see Appendix F2).

The number of students who participated and gave consent for their data to be used in this research was 237 students. Students' responses were quantitatively analyzed. Item difficulty, item discrimination, and percent response for each alternative were computed and used to better understand the characteristics of each item. Because there were more items (52 items) in the third draft of the instrument than the initial goal (40 items), item information was also used to decide which items to delete. Modifications were done to the assessment based on this psychometric analysis and the final version of the REALI instrument was obtained (see Appendix C3). Once more, all modifications were approved and supervised by two University of Minnesota faculty (Dr. Joan Garfield and Dr. Andrew Zieffler).

3.3.6 Field Test. In the development process, the pilot study was followed by the field test. Instructors who were currently teaching an introductory statistics course (undergraduate and graduate level) from colleges (2 and 4 years) and universities in the United States were invited to administer the REALI assessment to their students. Recruitment of these instructors was done by email through three listserv. An invitation letter was sent to (1) the *Consortium for the Advancement of Undergraduate Statistics Education (CAUSE)* website (http://www.causeweb.org); (2) the

statistics education section of the *American Statistical Association*; and (3) the *Isolated Statisticians* listserv (http://www2.lawrence.edu/fast/jordanj/isostat.html).

The invitation letter contained information about the purpose of the study and about the REALI instrument. To encourage participation, the invitation letter also provided incentives for instructors to participate. The incentives were that the instructors would receive a report, after their students completed the assessment, with information about what knowledge their students had regarding statistical literacy and reasoning and how they compared to students at other institutions (see Appendix G1).

The instructors who were interested in participating received a second email with additional information about the assessment. In addition, the instructors were asked to provide the following information: (1) institution name, (2) course name, (3) number of sections, (4) number of students in each section, and (5) short description of the curriculum (see Appendix G1). A follow up email was then sent with a copy of the REALI assessment, the link to the website containing the assessment, and a code to identify their students (see Appendix G1).

The administration of the instrument was done online through Qualtrics. The initial page of the assessment contained a consent form (see Appendix G2) asking if students were willing to participate in the study. A total of 758 students from introductory statistical courses representing 16 colleges and university took the REALI assessment. The following section reports the data analysis of the field test student's data.

3.4 Data Analysis

The analyses that will be used to answer the three research questions are described in more detail in the following sections. However, first a description of the variables used in this study is presented.

3.4.1 Variables. The *response variable* in this study is the students' responses to the items from the REALI assessment. All items in this assessment are multiple-choice; therefore, all students' answers will be dichotomous (1 for a correct answer and 0 for an incorrect answer). The *explanatory variables* are statistical literacy ability and statistical reasoning ability which are latent variables.

3.4.2 Research Question 1. The first research question to be answered is the following: What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony? To investigate the structure of the constructs, Item Response Theory (IRT), a common framework used in the measurement field, will be used. Five IRT models will be fitted to students' responses from the field test: one unidimensional IRT model, three bi-dimensional IRT models, and a bi-factor IRT model. For all IRT models the origin (mean of ability values) will be fixed to 0 and the unit (variance of ability values) will be fixed to 1.

The unidimensional IRT model will have all 40 items loading on only one dimension named Statistical Knowledge (see Figure 3.2). If the unidimensional model fits appropriately to the data, then this would provide some support to the assumption that there might not be a hierarchy between the constructs of statistical literacy and statistical reasoning. Instead, it could be argued that these two learning goals cannot be differentiated and they form a unique construct of statistical knowledge.

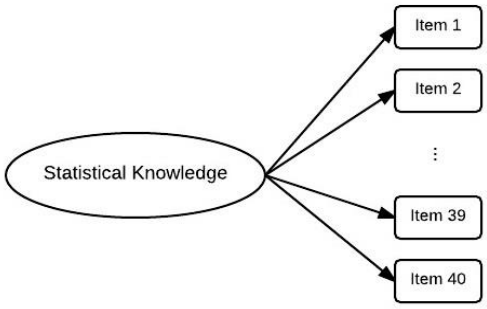


Figure 3.2. Unidimensional IRT model.

The Unidimensional Model will be modelled by the two parameter-logistic (2PL) IRT model. The 2PL model assumes unidimensionality of the test's scores. This model specifies the probability of a correct response to an item as a logistic function in which items are allowed to vary in terms of their difficulty and discrimination. In the 2PL model, the probability of a correct response is given by

$$p\left(x_{ij}=1|\,\theta,\alpha_j,\delta_j\right)=\frac{e^{\alpha_j\left(\theta-\delta_j\right)}}{1+e^{\alpha_j\left(\theta-\delta_j\right)}}, \tag{3.1}$$

where θ is the person-ability (or person-score), α_j is the discrimination parameter for item *j*, and δ_j is the difficulty for item *j*.

The remaining four models will consider a multidimensional structure between the learning goals of statistical literacy and statistical reasoning. Three models are bi-dimensional and one is a bi-factor model. The bi-factor model assumes a possible hierarchy between constructs. Thus statistical literacy and reasoning could be considered two sub-dimensions or subscores of a general construct of statistical knowledge.

Figure 3.3 shows each of the three bi-dimensional models. The Uncorrelated Model (Figure 3.3a) is composed of two uncorrelated dimensions: a statistical literacy dimension on which 20 items are loaded and a statistical reasoning dimension on which the remaining 20 items are loaded.

The Correlated Model (Figure 3.3b) is very similar to the Uncorrelated Model except that the two dimensions are now allowed to correlate. The third model is a Cross-loading Model which has the same structure of the Uncorrelated Model but there is a direct effect of the statistical literacy dimension on the statistical reasoning items. This can be seen in Figure 3.3c by the dotted lines which represent the cross-loadings. The cross-loadings will be fixed to be the same for all items. In addition, each of the statistical reasoning items will have its highest loading on the statistical reasoning dimension. Therefore, the cross-loadings will be fixed to a specific value that is smaller than the loadings from all the statistical reasoning items.

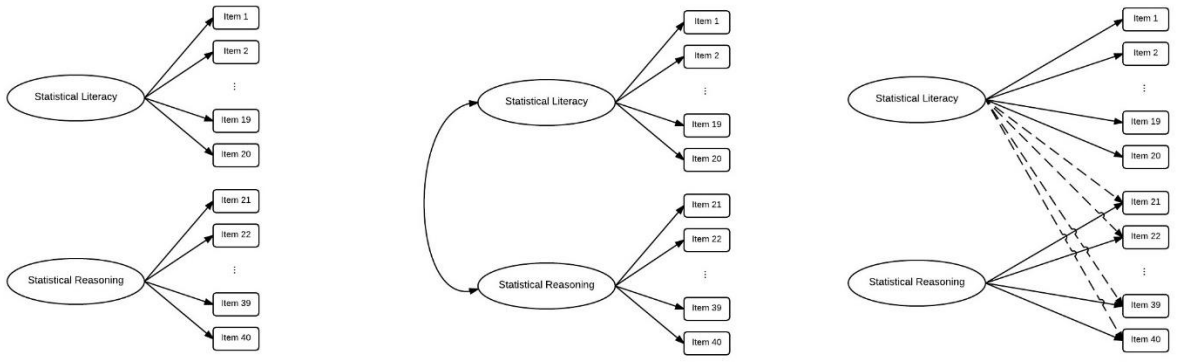


Figure 3.3. Bi-dimensional IRT models.

The three bi-dimensional models (Uncorrelated Model, Correlated Model, and Cross-loading Model) will be modelled using a multidimensional extension of the 2PL Model (McKinley & Reckase, 1983a; Reckase, 1985), where the probability of a response on item *j* by person *i* is given by

$$p\left(x_{ij}=1 \mid \boldsymbol{\theta}_{\mathbf{i}}, \boldsymbol{\alpha}_{\mathbf{j}}, \gamma_j\right)=\frac{e^{\left(\alpha_{j1}\theta_{i1}+\alpha_{j2}\theta_{i2}\right)+\gamma_j}}{1+e^{\left(\alpha_{j1}\theta_{i1}+\alpha_{j2}\theta_{i2}\right)+\gamma_j}}, \qquad (3.2)$$

where the intercept parameter is γ_j. The vector $\boldsymbol{\theta}_i$ corresponds to the ability (or subscore) of person *i* on each of the two dimensions. The vector $\boldsymbol{\alpha}_j$ contains the discrimination parameter of item *j* on each of the two dimensions, namely $\boldsymbol{\alpha}_j = (\alpha_{j1}, \alpha_{j2})$.

The final model that will be used to investigate the structure between statistical literacy and statistical reasoning is the Bi-factor Model A. Model Bi-factor A (Gibbons & Hedeker, 1992) has three dimensions: a general dimension (statistical knowledge) and two sub-dimensions (statistical literacy and statistical reasoning) (see Figure 3.4). This model is another multidimensional extension of the 2PL Model, where the probability of a response on item *j* by person *i* is given by

$$p\left(x_{ij} = 1 \mid \boldsymbol{\theta}_\mathbf{i}, \boldsymbol{\alpha}_\mathbf{j}, \gamma_j\right) = \frac{e^{\alpha_{jg}\theta_{ig} + \alpha_{js}\theta_{is} + \gamma_j}}{1 + e^{\alpha_{jg}\theta_{ig} + \alpha_{js}\theta_{is} + \gamma_j}}, \tag{3.3}$$

where θ_{ig} is the person-ability on the general factor, θ_{is} is the person-ability for each sub-dimension, and γ_j is a scalar parameter related to an overall multidimensional item difficulty. The vector $\boldsymbol{\alpha}_j$ contains the discrimination parameter of item *j* on the general dimension and on one of the two sub-dimensions (statistical literacy and statistical reasoning).

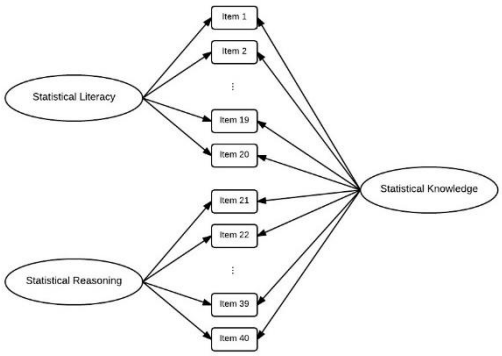


Figure 3.4. Bi-factor Model A.

Fit measures and model comparisons. The fit of the five IRT models to the REALI data will be evaluated at both the item- and the model-level. The S-X^2 is an item fit statistic (Orlando & Thissen, 2000, 2003) which will be used to assess if each item fits the IRT model. This statistic is based on the observed and expected frequencies correct and incorrect for each summed score. Under the hypothesis that the model fits the data and the sample size is large, the S-X^2 statistic is approximately distributed as a Pearson chi-squared statistic. Significant values indicate lack of fit.

The M_2 statistic, its associated *p*-value, and RMSEA (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005, 2006) will be used to evaluate model fit. The M_2 is a fit statistic that tests whether the model fits the data when all the items are considered simultaneously rather than individually. This statistic is based on the one- and two-way marginal tables of the complete cross-classification of the respondents based on their response patterns. Significant values indicate lack of fit. This statistic will be computed for each model. Models with non-significant results are preferred.

The Root Mean Square Error of Approximation (RMSEA) also gives information regarding the fit of the models. Guidelines suggested by Browne and Cudeck (1993) regarding RMSEA recommend that values ranging from 0.00 to 0.05 indicate close fit, values between 0.05 and 0.08 indicate fair fit, values ranging from 0.08 to 0.10 indicate mediocre fit, and values above 0.10 indicate unacceptable fit.

The IRT models will also be compared based on the likelihood ratio test (LRT; de Ayala, 2009), and two fit statistics: (1) the Akaike Information Criterion (AIC; Akaike, 1974) and (2) the Bayesian Information Criterion (BIC; Schwarz, 1978).

The deviance statistic will be computed for each of the three IRT models. Nested models will be compared using the likelihood ratio test (LRT; De Ayala, 2009). The LRT can be conducted by calculating the difference between two deviance statistics,

$$\Delta G^2 = -2\ln\left(L_{\text{reduced model}}\right) - \left[-2\ln\left(L_{\text{full model}}\right)\right] \text{ and} \quad (3.4)$$

$$\Delta G^2 = Deviance\ {}^{2}_{R} - Deviance\ {}^{2}_{F}. \quad (3.5)$$

If the sample size is large and the full model holds for the data, the likelihood ratio (ΔG^2) is distributed as chi-squared. Its degrees of freedom are given by the difference between the degrees of freedom from the reduced model and the degrees of freedom from the fuller model. Significant values indicate that the additional complexity of the full model provides more adequate fit.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) statistics from each model can also be compared. These measures are calculated as,

$$\text{AIC} = -2\text{lnL} + 2 * \text{P and} \quad (3.6)$$

$$\text{BIC} = -2\text{lnL} + \ln(\text{N}) * \text{P}, \quad (3.7)$$

respectively, where *P* is the number of parameters being estimated and *N* is the sample size. Smaller values of AIC and BIC indicate better fit.

The five IRT models will be compared based on the statistics reported above and the best fitting models will be chosen to answer the first research question. The IRT subscores from these models will be used to answer the second research question.

3.4.3 Research Question 2. The second research question to be answered is the following: What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction? The subscores from the best IRT models will be evaluated on distinctiveness and reliability. The Haberman analysis (Haberman, 2008) will also be used to evaluate the distinctiveness of subscores.

It is important to note that the dimensions of statistical literacy and statistical reasoning are most likely highly correlated. This would lead to highly correlated subscores of statistical literacy

and reasoning. However, for purposes of measurement, these dimensions need to be both reliable and distinct. To evaluate how reliable each of the subscores are, reliability estimates will be obtained for each IRT model using Equation 3.8. Models with higher reliability will be preferred.

$$\rho_S^2 = \frac{Var(\theta_S)}{Var(\theta_S) + MSE} \ , \tag{3.8}$$

where θ_s is the person-ability for each dimension and MSE is the mean of the conditional error variance for the θ_s estimates for all students.

To address the distinctiveness issue, the IRT models will be compared in order to find the model that decreases the correlation between subscores of statistical literacy and reasoning. Reliability of the differences between person subscores will also be computed and compared between models. Models with higher values for the reliability of the difference have more evidence of distinction and will be preferred. The following equation will be used to compute the reliability of the difference between person's subscores.

$$\rho^2_{S1-S2} = \frac{\rho^2_{S1}\sigma^2_{\theta_{S1}} + \rho^2_{S2}\sigma^2_{\theta_{S2}} - 2cor(S1,S2)\sigma_{\theta_{S1}}\sigma_{\theta_{S2}}}{\sigma^2_{\theta_{S1}} + \sigma^2_{\theta_{S2}} - 2cor(S1,S2)\sigma_{\theta_{S1}}\sigma_{\theta_{S2}}}, \tag{3.9}$$

where the reliability of each subscore is represented by ρ^2_{S1} and ρ^2_{S2}, $cor(S1,S2)$ represents the correlation between the susbscores, and $\sigma^2_{\theta_{S1}}$ and $\sigma^2_{\theta_{S1}}$ represents the variance of each subscore.

In addition, the Haberman analysis (Haberman, 2008) will also be conducted to verify if reporting two subscores (statistical reasoning and statistical literacy) is more valuable than reporting a total score. Haberman (2008) suggests different ways to estimate the true subscore: (1) using a linear regression of the true subscore (τ_x) on the observed subscore (S_x) and (2) using a linear regression of the true subscore (τ_x) on the observed total score (S_z). This analysis aims to verify if one of these approximations of the true subscore is better or have added value over the

other. The proportional reduction in mean square error (PRMSE) will be used to verify which approximation of the true subscore is more accurate. The approximation with the smallest mean square error (MSE), and therefore the one with the greatest value of PRMSE, is more accurate at estimating the true subscore. The literature does not contain a clear guideline for how much greater the PRMSE(S_x) needs to be in relation to PRMSE(S_z) to state that the subscores have added value over the total score. However, a difference in PRMSE values of 0.01 has been used in the literature as enough evidence to state the usefulness of subscores over the total score (Sinharay, 2010). The reliability of each IRT subscore will be computed using the Equation 3.8. The equations to estimate the MSE and the PRMSE for the two approximations of the true subscore (τ_x) are presented in Table 3.6.

Table 3.6

Equations for the estimation of MSE and PRMSE.

	Predictor	
	Observed subscore (S_x)	Observed total score (S_z)
MSE	$\sigma^2(\tau_x)[1-\rho^2(S_x,\tau_x)]$	$\sigma^2(\tau_x)[1-\rho^2(S_z,\tau_x)]$
PRMSE	$\rho^2(S_x,\tau_x)$	$\rho^2(S_z,\tau_x)$

3.4.4 Research Question 3. The third research question to be answered is the following: What measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning? To answer this third research question, the results from the previous two research questions will be considered and a final model will be chosen. To choose this final model, careful attention will be given to finding a model that not only provides evidence of good fit (results of research question 1) but also provides evidence supporting the distinction and reliability of the subscores (results of research question 2).

3.5 Chapter Summary

This chapter presented the development process of the REALI assessment which included development of the test blueprint, expert review, think-aloud interview, pilot test, and field test. The chapter also described the data collection and data analysis. The next chapter presents the results of the study.

Chapter 4

Results

This chapter reports the results from the development process of the REALI assessment and the results and analysis of the pilot and field test. A summary of expert feedback is provided, followed by the results from the think-aloud interviews with students. Next, students' responses from the pilot test are presented. Finally, student's data from the field test are also reported along with the results from psychometric and subscore analysis used to answer the research questions posed by this study.

4.1 Expert Review Feedback

Four professionals in statistics education were identified and requested to serve as expert reviewers of the REALI instrument. The review consisted of categorizing each of the 52 items into two groups. In addition, the reviewers were also invited to critique and provide feedback about the items. Two experts were able to complete the review on time and two experts submitted the review more than two months after the due date. Because of the timeline of the study and the need to reach a final version of the assessment, the changes done to the items were only based on the review from the two experts who submitted the review on time. However, the categorization of items from all four experts was used in the study because this categorization was not an essential process needed to reach a final version of the assessment.

4.1.1 Categorization of items. Appendix D4 shows the categorization of each of the items by each of the four experts. Experts were requested to rate all 52 items; however, one of the experts only had time to complete the categorizations for 34 items. In addition, some experts reported being unsure of the categorization of some of the items (these ratings were not considered in this analysis).

Out of the 52 items, 27 items had ratings from all four experts, 23 items had ratings from three experts, and one item had ratings from two experts (see last row in Table 4.1). One of the items (Item 8D) was rated by only one expert; therefore, this item was not included in this analysis. After the categorization was done, the percentage of experts whose categorization was the same as that determined by this researcher, was calculated for each item. For instance, Table 4.2 shows the categorization of Item 1A. This item had ratings from each of the experts, being a total of four ratings. Out of the four ratings, three agreed with the categorization of the author. Therefore, the percentage of agreement was 3/4 = 0.75 or 75%. Table 4.1 shows the percentage of experts who categorized the items in the same way as the author and the number of items for each percentage.

Table 4.1

Number of items for each rating and for each percentage of expert agreement.

Percentage	Number of Items			Total
	4 ratings	3 ratings	2 ratings	
100%	7	10	-	17
75%	12	-	-	12
67%	-	6	-	6
50%	6	-	1	7
33%	-	4	-	4
25%	1	-	-	1
0%	1	3	-	4
Total	27	23	1	51

Out of the 52 items, 35 items (17 + 12 + 6) had more than 50% of the experts agreeing with the categorization of the author. For 7 items, exactly half of the experts agreed with the author's categorization, and 9 items (4 + 1 + 4) had less than 50% of the experts agreeing with the categorization of the author.

Table 4.2

Example of categorization and percentage agreement for Item 1A.

Item	Categorization					Percentage
	Author	Expert A	Expert B	Expert C	Expert D	
1A	Statistical Literacy	Statistical Reasoning	Statistical Literacy	Statistical Literacy	Statistical Literacy	75%

Table 4.3 and Figure 4.1 show each percentage of agreement (0%, 25% … and 100%) and how many items were rated for percentage of agreement. In general, there were more statistical literacy items that received 100% or 75% of agreement in the categorization by the experts, and fewer statistical reasoning items had as much agreement in their categorization. These results suggest that it was easier for the experts to agree on the categorization of statistical literacy items than to agree on the categorization of statistical reasoning items.

Table 4.3

Number and percentage of items for each percentage of expert agreement categorized by learning goals.

Percentage	S. Literacy	S. Reasoning	Total
100%	14 (50%)	3 (13%)	17
75%	8 (29%)	4 (17%)	12
67%	3 (11%)	3 (13%)	6
50%	2 (7%)	5 (22%)	7
33%	1 (4%)	3 (13%)	4
25%	0 (0%)	1 (4%)	1
0%	0 (0%)	4 (17%)	4
Total	28 (100%)	23 (100%)	51

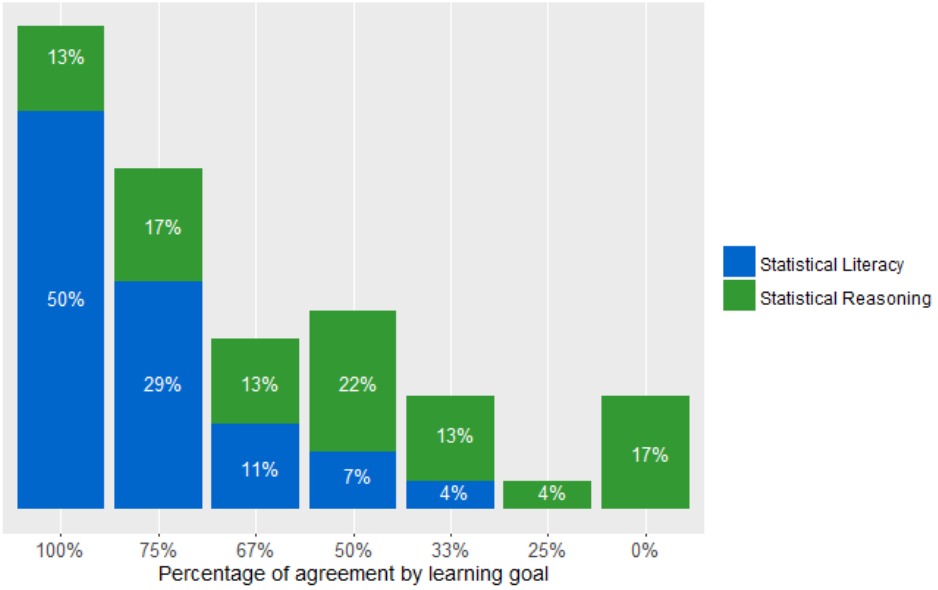


Figure 4.1. Percentage of statistical literacy and reasoning items for each percentage agreement category.

4.1.2 Item changes. In addition to categorizing the items, the experts were also invited to critique them. As a result, changes were made to the items. All changes are shown in Appendix I1. Items were changed, based on expert feedback, to increase clarity, decrease the length of the correct alternatives, delete weak alternatives, and to ensure that alternatives were balanced.

After the items were changed based on expert review, the author looked again at all items to check for additional problems. Item 2B was the only item with concerns. Changes made to this item are also shown in Appendix I1.

4.2 Think-Aloud Interviews

Think-aloud interviews were conducted with students to provide evidence if students were responding to the items as intended. In addition, these interviews were used to gather evidence of response process to clarify the categorization of some items. These interviews were conducted with four students from the University of Minnesota. One interviewee was an undergraduate student

from the Statistics Department enrolled in a second course of statistics. The other three interviewees were master's students: two were enrolled in a second course of statistics in the Educational Psychology Department and the other student had taken an introductory statistics course in the previous semester in the same department. Two out of the four students were able to complete the whole assessment during the one hour think-aloud interview.

Students' responses from the think-aloud interview were used to make improvements to the REALI assessment. Major changes done to the items and major concerns found during the think-aloud interviews are shown in Appendix I2.

After the items were changed based on the think-aloud interviews, a graduate student at the University of Minnesota was asked to read each item from the REALI assessment looking for any additional problems. Based on her feedback, small changes were done to some of the items to improve clarity. These changes are shown in Appendix I1.

4.3 Pilot Test

The third draft of the REALI instrument was used in the pilot test. A total of 237 students completed the assessment and gave consent for the author to use their data in this research. The purpose of the pilot test was to estimate item difficulty and item discrimination and to verify if there was any item not behaving properly.

All 237 students were enrolled in introductory statistics courses at the undergraduate or beginning graduate level. Two hundred and four students were from the University of Minnesota and 33 students were from Augsburg College. Out of the 204 University of Minnesota students, 96 were from an undergraduate-level statistics course, 69 students were from a graduate-level statistics course, and 39 students were from a bio-statistics graduate level course. The 33 students from Augsburg College were from four undergraduate-level statistics courses.

An examination of item characteristics for each of the 48 items on the REALI instrument is presented below. All of the computations were conducted using R version 3.3.0 (R Development Core Team, 2016)

Table 4.4 shows the proportion of students that chose each of the alternatives for each of the 48 items. Item difficulty (proportion correct) and item discrimination (point biserial correlation corrected for spuriousness) were also computed for each of the REALI items under the Classical Test Theory framework. These values are presented in Table 4.5. Item difficulty values ranged from 0.19 (very difficult items) to 0.95 (very easy items). Item discrimination values ranged from 0.01 to 0.50. Items with discrimination values lower than 0.20 are usually considered poorly discriminating items. A total of six items were flagged (items 1, 4, 36, 40, 44, and 48) as poorly discriminating items.

Table 4.4

Proportion correct for each alternative for each item.

Item	A	B	C	D	E	Item	A	B	C	D	E
1	0.12	0.46	0.42*			25	0.14	0.07	0.74*	0.06	
2	0.16	0.09	0.75*			26	0.73*	0.21	0.06		
3	0.07	0.52	0.36*	0.05		27	0.41	0.47*	0.11		
4	0.24	0.55*	0.21			28	0.10	0.05	0.80*	0.05	
5	0.11	0.08	0.02	0.79*		29	0.06	0.15	0.34	0.45*	
6	0.04	0.71*	0.25			30	0.75*	0.13	0.12		
7	0.57*	0.21	0.09	0.13		31	0.13	0.14	0.60*	0.12	
8	0.11	0.14	0.75*			32	0.13	0.59*	0.16	0.12	
9	0.21	0.51	0.58*			33	0.16	0.10	0.16	0.59*	
10	0.03	0.07	0.08	0.81*		34	0.27	0.67*	0.06		
11	0.44*	0.14	0.21	0.21		35	0.25	0.26	0.44*	0.05	
12	0.11	0.81*	0.08			36	0.19*	0.05	0.35	0.40	
13	0.50*	0.32	0.19			37	0.19	0.29	0.52*		
14	0.26	0.71*	0.03			38	0.17	0.14	0.19	0.51*	
15	0.10	0.10	0.80*			39	0.07	0.55*	0.38		
16	0.32	0.11	0.20	0.36*		40	0.15	0.18	0.11	0.56*	
17	0.09	0.09	0.82*			41	0.04	0.11	0.05	0.17	0.62*
18	0.57*	0.21	0.22			42	0.22	0.40*	0.38		

Item	A	B	C	D	E	Item	A	B	C	D	E
19	0.14	0.49*	0.11	0.26		43	0.74*	0.10	0.16		
20	0.08	0.08	0.04	0.80*		44	0.28*	0.19	0.53		
21	0.14	0.51*	0.30	0.05		45	0.95*	0.04	0.01		
22	0.08	0.08	0.84*			46	0.18	0.11	0.65*	0.06	
23	0.16	0.79*	0.05			47	0.34*	0.42	0.24		
24	0.62*	0.22	0.16			48	0.18	0.45	0.29*	0.08	

Note. Items with no results presented for Option D (or Option E) represent an item that did not have an Option D (or Option E). * indicates correct answer.

Because the version of the REALI assessment used in the pilot test had more items than needed, it was necessary to decide which items would be deleted for each area of learning. Appendix I3 contains information about which items were deleted and about changes done to the items after the pilot study.

Table 4.5

Item difficulty, item discrimination, areas of learning and learning goal for each item.

Item	Difficulty	Discrimination	Learning Goal	Area of Learning
1	0.42	0.12*	Literacy	Representations of data
2	0.75	0.27	Literacy	
3	0.36	0.33	Reasoning	
4	0.55	0.18*	Reasoning	
5	0.79	0.40	Reasoning	
6	0.71	0.25	Literacy	Measures of center
7	0.57	0.30	Literacy	
8	0.75	0.42	Reasoning	
9	0.28	0.27	Reasoning	
10	0.81	0.41	Literacy	Measures of variability
11	0.44	0.36	Literacy	
12	0.81	0.24	Reasoning	
13	0.50	0.30	Reasoning	
14	0.71	0.39	Literacy	Study design
15	0.80	0.49	Literacy	
16	0.36	0.32	Literacy	
17	0.82	0.44	Reasoning	
18	0.57	0.28	Literacy	
19	0.49	0.28	Reasoning	
20	0.80	0.45	Reasoning	

Item	Difficulty	Discrimination	Learning Goal	Area of Learning
21	0.51	0.37	Reasoning	
22	0.84	0.42	Literacy	Hypothesis testing and p-value
23	0.79	0.34	Literacy	
24	0.62	0.35	Reasoning	
25	0.74	0.50	Literacy	
26	0.73	0.39	Literacy	
27	0.47	0.20	Reasoning	
28	0.80	0.46	Reasoning	
29	0.45	0.33	Literacy	
30	0.75	0.39	Reasoning	
31	0.60	0.48	Literacy	
32	0.59	0.36	Literacy	
33	0.59	0.38	Reasoning	
34	0.67	0.30	Reasoning	
35	0.44	0.42	Literacy	Confidence interval
36	0.19	0.04*	Literacy	
37	0.52	0.37	Reasoning	
38	0.51	0.37	Reasoning	
39	0.55	0.21	Reasoning	Probability
40	0.56	0.06*	Literacy	
41	0.62	0.35	Literacy	
42	0.40	0.23	Reasoning	
43	0.74	0.27	Literacy	
44	0.28	0.01*	Reasoning	
45	0.95	0.26	Literacy	Bivariate data
46	0.65	0.40	Literacy	
47	0.34	0.23	Reasoning	
48	0.29	0.14*	Reasoning	

Note. * indicates items with discrimination values smaller than 0.2.

4.4 Field Test

This subsection reports the results from the field test using students from introductory statistics courses. A descriptive analysis is provided followed by the results of fitting the five different IRT models. Then, the model comparison is reported followed by the results of subcore correlation, subscore reliability, and the Haberman analysis.

4.4.1 Descriptive Analysis. The field test was performed at the end of the spring semester during the months of April and May. A total of 758 students from introductory statistics courses took the REALI assessment. However, the sample used in the data analysis was composed only of students who had consented to participate in the study (96% of the students), responded to all 40 items, and had completed the assessment in at least 10 minutes. The time constraint criterion was used to eliminate students who did not engage sufficiently with the test items. The final sample was composed of 671 students representing 16 universities and colleges around the United States and Canada.

Summaries of the total raw score and raw subscores, including the computation of coefficient alpha, are provided next together with correlations between total raw score and raw subscores. In addition, the proportion of students who selected each of the item answer choices are provide below as well as an examination of item difficulty and item discrimination for each of the 40 items. All analyses were conducted using R version 3.3.0 (R Development Core Team, 2016).

A histogram of the distribution of the REALI total raw scores for the 671 students in the sample is presented in Figure 4.2. The mean of these scores was 24.16 (*SD* = 7.48) and the median was 24. The minimum and maximum values observed were 4 and 40 respectively. The estimate of the internal consistency, coefficient alpha, for these scores was 0.87.

The statistical literacy and statistical reasoning raw subscores were investigated to better understand how students were performing in statistical literacy and statistical reasoning items. Histograms of the distributions of the two raw subscores for the 671 students in the sample are presented in Figure 4.3. The mean statistical literacy subscore was 13.15 (*SD* = 3.82) and the mean statistical reasoning subscore was 11.01 (SD = 4.15). Table 4.6 shows other descriptive statistics for each of the subscores. The estimate of the internal consistency, coefficient alpha, for the statistical literacy subscore was 0.76 and for the statistical reasoning subscore was 0.78.

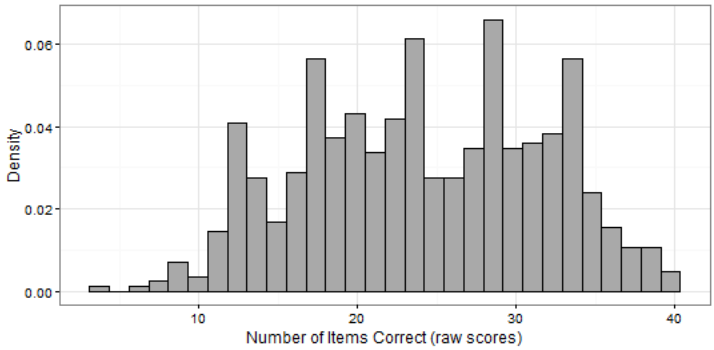


Figure 4.2. Distribution of total scores.

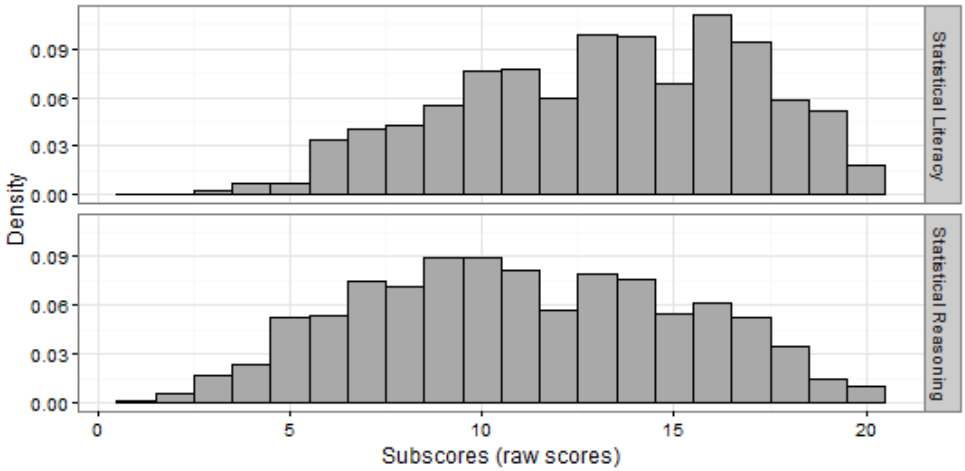


Figure 4.3. Distribution of the statistical literacy and statistical reasoning subscores.

Table 4.6

Descriptive statistics for each of the subscores

Summary Measure	Statistical Literacy	Statistical Reasoning
Mean	13.15	11.01
SD	3.82	4.15
Median	13	11
Minimum	3	1
Maximum	20	20

In order to examine the relationship between the total raw scores and the statistical reasoning and statistical literacy subscores, correlations and scatterplots are presented. Figure 4.4 shows the scatterplots of the two subscores. The statistical literacy and statistical reasoning

subscores had a correlation of 0.76. Figure 4.5 shows the scatterplots between the total score and the statistical literacy and statistical reasoning subscores. The correlation between the total score and the statistical literacy and statistical reasoning subscores were very high, 0.93 and 0.94 respectively.

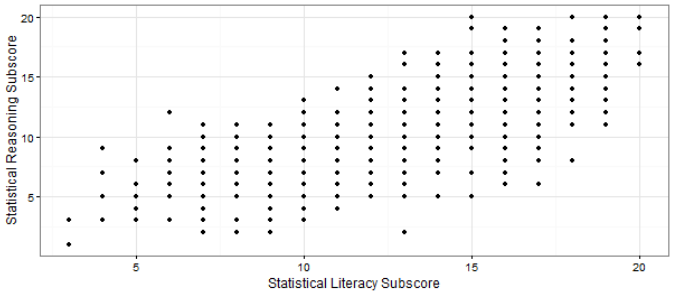


Figure 4.4. Scatterplots of the statistical literacy and statistical reasoning subscores.

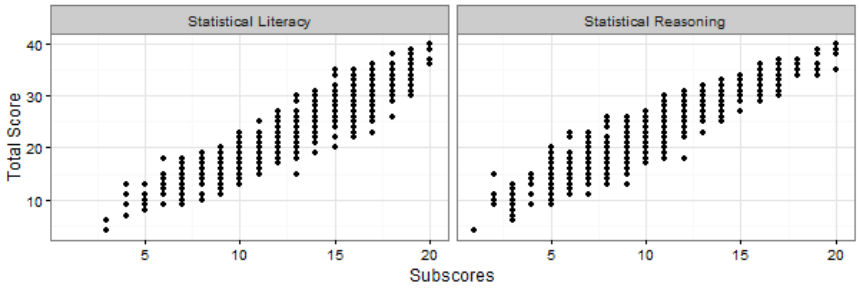


Figure 4.5. Scatterplots of the total score and the statistical literacy and statistical reasoning subscores.

To better understand how each distractor (incorrect alternative options) was behaving, the proportion of students that chose each of the alternatives for each of the 40 items was computed (see Table 4.7). In addition, item discrimination (point biserial correlation corrected for spuriousness) and item difficulty (proportion correct) were computed for each of the REALI items. These values are presented in Table 4.8. Item difficulty can be used to better understand how difficult or easy each item is. Item discrimination gives information about how good each item is at discriminating between high and low ability students.

Table 4.7

Proportion of students who chose each of the alternatives for all 40 items

Item	A	B	C	D	E	Item	A	B	C	D	E
1	0.07	0.09	0.09	0.75*		21	0.09	0.10	0.67*	0.14	
2	0.17	0.10	0.72*			22	0.77*	0.16	0.07	-	
3	0.06	0.56	0.29*	0.09		23	0.12	0.10	0.73*	0.06	
4	0.08	0.07	0.05	0.80*		24	0.59*	0.19	0.21		
5	0.02	0.97*	0.01			25	0.14	0.18	0.56*	0.12	
6	0.53*	0.26	0.07	0.14		26	0.17	0.51*	0.12	0.21	
7	0.13	0.21	0.66*			27	0.21	0.13	0.17	0.49*	
8	0.19	0.54	0.27*			28	0.26	0.64*	0.11		
9	0.07	0.03	0.10	0.80*		29	0.22	0.28	0.41*	0.09	
10	0.45*	0.15	0.27	0.14		30	0.57*	0.26	0.11	0.06	
11	0.14	0.78*	0.08			31	0.11	0.38	0.51*		
12	0.47*	0.31	0.22			32	0.16	0.18	0.20	0.46*	
13	0.28	0.70*	0.02			33	0.10	0.60*	0.31		
14	0.15	0.08	0.77*			34	0.06	0.09	0.07	0.24	0.54*
15	0.28	0.13	0.18	0.41*		35	0.29	0.46*	0.25		
16	0.10	0.11	0.79*			36	0.73*	0.09	0.17		
17	0.21	0.40*	0.19	0.20		37	0.93*	0.05	0.01		
18	0.12	0.09	0.04	0.74*		38	0.19	0.13	0.65*	0.03	
19	0.10	0.19	0.71*			39	0.40*	0.38	0.21		
20	0.58*	0.27	0.15			40	0.14	0.19	0.11	0.34*	0.22

Note. * indicates correct answer.

Table 4.8

Item difficulty and item discrimination for all 40 items

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
1	0.75	0.26	21	0.67	0.52
2	0.72	0.36	22	0.77	0.31
3	0.29	0.28	23	0.73	0.48
4	0.80	0.43	24	0.59	0.32
5	**0.97**	**0.05**	25	0.56	0.44
6	**0.53**	**0.15**	26	0.51	0.42
7	0.66	0.36	27	0.49	0.34
8	0.27	0.31	28	0.64	0.26
9	0.80	0.40	29	0.41	0.38
10	0.45	0.45	**30**	**0.57**	**0.15**
11	0.78	0.26	31	0.51	0.50

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
12	0.47	0.34	32	0.46	0.43
13	0.70	0.37	33	0.60	0.22
14	0.77	0.37	34	0.54	0.47
15	0.41	0.35	35	0.46	0.35
16	0.79	0.44	**36**	**0.73**	**0.18**
17	0.40	0.36	37	0.93	0.25
18	0.74	0.45	38	0.65	0.43
19	0.71	0.49	39	0.40	0.35
20	0.58	0.45	40	0.34	0.36

Note. Bold font indicates items with discrimination values less than 0.2.

The distribution of discrimination values (Figure 4.6) shows all positive values of discrimination, ranging from 0.05 to 0.52. Values of discrimination higher than 0.2 are desired. Only four out of the 40 items presented low discrimination. These are items 5, 6, 30, and 36 which presented discrimination values 0.05, 0.15, 0.15, and 0.18, respectively. The distribution of item difficulty values shown in Figure 4.7 indicates much variation with difficulties ranging from 0.27 (very difficult items) to 0.97 (very easy items).

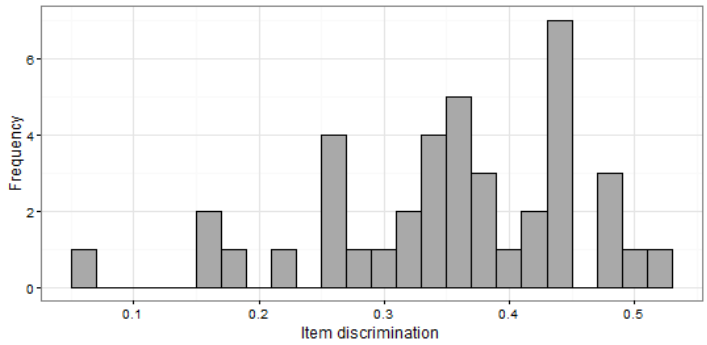


Figure 4.6. Distribution of item discrimination values.

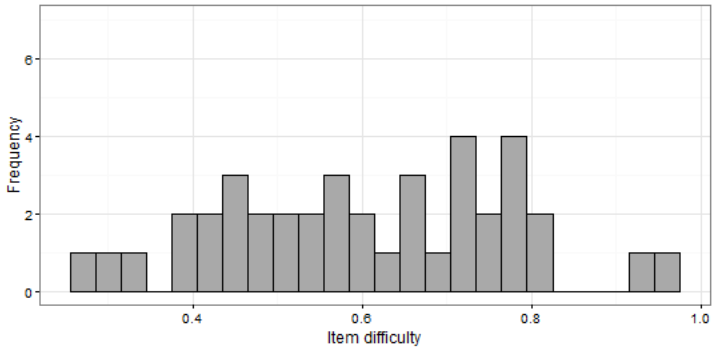


Figure 4.7. Distribution of item difficulty values.

4.4.2 Item Response Theory Models. Five IRT models were fitted to the REALI data to better understand what measurement model best represents the construct of statistical literacy and the construct of statistical reasoning. One unidimensional model, three bi-dimensional, and a bi-factor model: (1) Unidimensional Model, (2) Uncorrelated Model, (3) Correlated Model, (4) Cross-loading Model, and (5) Bi-factor Model A. Item parameters, item-level diagnostics statistics, and fit statistics are reported for each of the five IRT models in the next subsection.

Unidimensional Model. Table 4.9 provides the item parameters (intercept, item discrimination, and item difficulty) and standard error of the estimates obtained from fitting the Unidimensional Model to the data. The estimated item discrimination parameters ranged from 0.32 to 1.78. Based on guidelines available in De Ayala (2009), nine of the 40 items on REALI presented discrimination values lower than 0.8 which indicates poor discrimination. Item difficulty values ranged from –2.42 to 1.37, except for Item 5 which presented an uncommon value for item difficulty (-10.51), due to a very high percentage of students who correctly responded to this item (97%).

Table 4.9

Item Parameters for the Unidimensional Model

	Discrimination		Intercept		Difficulty	
Item	*a*	*s.e.*	*c*	*s.e.*	*b*	*s.e.*
1*	0.70	0.11	1.20	0.10	-1.72	0.27
2	1.07	0.13	1.20	0.12	-1.12	0.13
3*	0.72	0.10	-0.98	0.10	1.37	0.21
4	1.66	0.19	2.04	0.19	-1.23	0.10
5*	0.33	0.25	3.48	0.23	-10.51	7.80
6*	0.35	0.09	0.12	0.08	-0.35	0.24
7	0.99	0.12	0.82	0.10	-0.83	0.12
8	0.84	0.11	-1.14	0.10	1.35	0.19
9	1.42	0.17	1.87	0.16	-1.32	0.12
10	1.24	0.12	-0.26	0.11	0.21	0.09
11*	0.78	0.12	1.41	0.11	-1.80	0.26
12	0.87	0.10	-0.14	0.09	0.16	0.11
13	1.09	0.13	1.04	0.12	-0.95	0.12
14	1.21	0.15	1.55	0.14	-1.28	0.13
15	0.91	0.11	-0.42	0.10	0.46	0.12
16	1.72	0.20	1.97	0.19	-1.15	0.09
17	0.89	0.11	-0.45	0.10	0.51	0.12
18	1.54	0.17	1.53	0.16	-0.99	0.09
19	1.71	0.18	1.41	0.16	-0.83	0.08
20	1.21	0.13	0.44	0.11	-0.37	0.09
21	1.79	0.19	1.16	0.16	-0.65	0.07
22	0.94	0.13	1.43	0.12	-1.53	0.19
23	1.65	0.18	1.48	0.16	-0.90	0.08
24	0.83	0.11	0.44	0.09	-0.53	0.12
25	1.24	0.13	0.32	0.11	-0.26	0.08
26	1.13	0.12	0.06	0.10	-0.05	0.09
27	0.80	0.10	-0.01	0.09	0.02	0.11
28*	0.64	0.10	0.62	0.09	-0.96	0.18
29	0.99	0.11	-0.40	0.10	0.40	0.11
30*	0.32	0.09	0.30	0.08	-0.94	0.34
31	1.44	0.14	0.10	0.11	-0.07	0.08
32	1.19	0.13	-0.19	0.10	0.16	0.09
33*	0.50	0.09	0.41	0.08	-0.83	0.21
34	1.35	0.14	0.24	0.11	-0.18	0.08
35	0.87	0.11	-0.18	0.09	0.21	0.11
36*	0.44	0.10	1.06	0.09	-2.42	0.56
37	1.45	0.28	3.40	0.31	-2.34	0.29
38	1.25	0.14	0.81	0.12	-0.65	0.09
39	0.87	0.11	-0.44	0.09	0.51	0.12
40	0.96	0.12	-0.79	0.10	0.82	0.13

Note. * indicates item with discrimination value lower than 0.8;

Item-level diagnostics statistics are presented in Table 4.10. A total of 10 items presented item misfit. Two additional fit indices were computed: the RMSEA was 0.03 and the M_2 (df = 740) was 1298.61 (p = .0001).

Table 4.10

Item level diagnostics statistics for the Unidimensional Model

Item	X^2	*d.f.*	Probability	Item	X^2	*d.f.*	Probability
1	32.31	26	0.1825	21	25.45	23	0.3266
2	**42.63**	**26**	**0.0211**	22	37.11	26	0.0728
3	35.01	27	0.1383	23	24.58	23	0.3742
4	**37.03**	**22**	**0.0234**	24	30.80	28	0.3254
5	5.61	12	0.9345	**25**	**55.76**	**25**	**0.0004**
6	29.86	29	0.4228	**26**	**40.70**	**27**	**0.0439**
7	**43.54**	**26**	**0.0169**	**27**	**44.32**	**27**	**0.0191**
8	38.10	27	0.0760	28	34.17	28	0.1949
9	**35.51**	**23**	**0.0461**	29	25.23	26	0.5076
10	23.96	26	0.5794	30	35.21	31	0.2748
11	29.74	27	0.3251	31	31.37	25	0.1765
12	37.96	27	0.0783	32	22.44	26	0.6651
13	36.86	26	0.0768	33	26.69	29	0.5896
14	26.35	24	0.3349	**34**	**43.48**	**25**	**0.0124**
15	25.89	27	0.5258	35	36.26	26	0.0868
16	15.69	22	0.8316	36	37.06	27	0.0938
17	31.80	27	0.2388	37	17.62	19	0.5495
18	**46.66**	**23**	**0.0025**	**38**	**44.45**	**25**	**0.0096**
19	29.79	23	0.1550	39	29.37	27	0.3423
20	21.81	25	0.6478	40	38.89	27	0.0647

Note. The bold font indicates *p*-value smaller than 0.05.

Uncorrelated Model. Table 4.11 provides the item parameters (intercept and discriminations) and standard error of the estimates obtained from fitting the Uncorrelated Model to the data. Item intercept values ranged from –1.15 to 3.47. The estimated item discrimination values for the statistical literacy dimension ranged from 0.25 to 1.78. Based on guidelines available in De Ayala (2009), five of the 20 literacy items on REALI presented discrimination values lower

than 0.8 which indicates poor discrimination. Item discrimination parameters for the statistical reasoning dimension ranged from 0.57 to 1.62. Based on guidelines from De Ayala (2009), five of the 20 reasoning items presented poor discrimination. The covariance between the statistical literacy and the statistical reasoning constructs were set to be zero.

Table 4.11

Item Parameters for the Uncorrelated Model

	Discrimination				Intercept	
	Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	*c*	*s.e.*
1*	0.65	0.12	0.00	-----	1.19	0.10
2	1.06	0.14	0.00	-----	1.19	0.11
3*	0.00	-----	0.70	0.11	-0.98	0.10
4	0.00	-----	1.52	0.20	1.95	0.17
5*	0.28	0.26	0.00	-----	3.47	0.23
6*	0.32	0.09	0.00	-----	0.12	0.08
7	0.00	-----	1.03	0.13	0.83	0.10
8	0.00	-----	0.87	0.12	-1.15	0.10
9	1.56	0.19	0.00	-----	1.96	0.17
10	1.02	0.12	0.00	-----	-0.26	0.09
11	0.00	-----	0.84	0.13	1.43	0.11
12	0.00	-----	0.94	0.12	-0.14	0.09
13	1.07	0.13	0.00	-----	1.02	0.11
14	1.12	0.15	0.00	-----	1.50	0.13
15	0.90	0.12	0.00	-----	-0.42	0.09
16	0.00	-----	1.62	0.21	1.90	0.17
17*	0.00	-----	0.78	0.11	-0.44	0.09
18	0.00	-----	1.56	0.19	1.53	0.15
19	1.66	0.19	0.00	-----	1.37	0.14
20	0.00	-----	1.26	0.14	0.44	0.10
21	1.78	0.20	0.00	-----	1.13	0.14
22	1.02	0.14	0.00	-----	1.47	0.12
23	0.00	-----	1.49	0.18	1.39	0.14
24*	0.00	-----	0.77	0.11	0.43	0.09
25	1.38	0.15	0.00	-----	0.33	0.10
26	1.14	0.13	0.00	-----	0.05	0.09
27	0.00	-----	0.80	0.11	-0.02	0.08
28*	0.00	-----	0.57	0.10	0.60	0.08

	Discrimination				Intercept	
	Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	*c*	*s.e.*
29	0.98	0.12	0.00	-----	-0.41	0.09
30*	0.25	0.09	0.00	-----	0.30	0.08
31	0.00	-----	1.45	0.16	0.10	0.10
32	0.00	-----	1.16	0.13	-0.19	0.09
33*	0.00	-----	0.61	0.10	0.43	0.08
34	1.37	0.15	0.00	-----	0.23	0.10
35	0.00	-----	1.04	0.12	-0.19	0.09
36*	0.50	0.11	0.00	-----	1.07	0.09
37	1.20	0.24	0.00	-----	3.21	0.26
38	1.21	0.14	0.00	-----	0.79	0.10
39	0.00	-----	0.90	0.12	-0.45	0.09
40	0.00	-----	1.03	0.13	-0.81	0.10

Note. * indicates item with discrimination value lower than 0.8;

Item-level diagnostics statistics are presented in Table 4.12. A total of nine items presented item misfit. The other fit indices were the RMSEA of 0.04 and the M_2 (*df* = 740) of 1671.27 (*p* = .0001).

Table 4.12

Item level diagnostics statistics for the Uncorrelated Model

Item	X^2	*d.f.*	Probability	Item	X^2	*d.f.*	Probability
1	33.00	28	0.2352	21	36.96	25	0.0581
2	**41.78**	**27**	**0.0345**	22	38.13	27	0.0756
3	34.32	27	0.1566	23	32.05	25	0.1561
4	**39.61**	**24**	**0.0235**	24	29.91	29	0.4202
5	7.65	12	0.8123	**25**	**63.03**	**27**	**0.0001**
6	31.22	30	0.4060	**26**	**47.10**	**27**	**0.0096**
7	**42.14**	**27**	**0.0318**	**27**	**44.00**	**27**	**0.0206**
8	37.25	28	0.1132	28	36.01	28	0.1419
9	33.35	25	0.1222	29	31.10	27	0.2665
10	40.09	27	0.0502	30	36.07	31	0.2426
11	29.47	27	0.3376	31	31.65	25	0.1679
12	37.75	28	0.1029	32	22.81	26	0.6449
13	39.55	27	0.0563	33	27.17	29	0.5636
14	30.05	25	0.2220	**34**	**49.88**	**26**	**0.0032**
15	30.39	28	0.3441	35	36.32	26	0.0858

Item	X^2	*d.f.*	Probability	Item	X^2	*d.f.*	Probability
16	18.15	24	0.7962	36	35.71	29	0.1816
17	33.22	27	0.1894	37	18.87	19	0.4666
18	**46.51**	**25**	**0.0056**	**38**	**46.48**	**27**	**0.0113**
19	35.68	25	0.0764	39	30.02	27	0.3124
20	23.25	25	0.5643	40	39.62	27	0.0554

Note. The bold font indicates *p*-value smaller than 0.05.

Figure 4.8 shows the scatterplot of the statistical literacy subscore and statistical reasoning subscore produced by the Uncorrelated Model. The correlation between these subscores was 0.684.

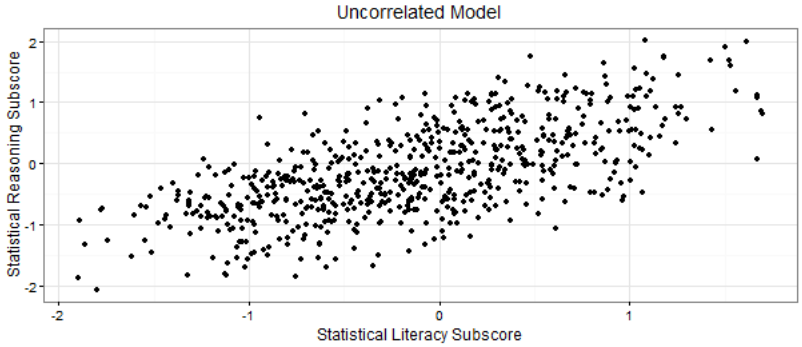


Figure 4.8. Scatterplots of the statistical literacy and statistical reasoning subscores for the Uncorrelated Model.

Correlated Model. Table 4.13 shows the item parameters (intercept and discriminations) and standard error of the estimates obtained from fitting the Correlated Model to the data. Item intercept values ranged from –1.14 to 3.48. The estimated item discrimination parameters for the statistical literacy dimension ranged from 0.32 to 1.79. Item discrimination parameters for the statistical reasoning dimension ranged from 0.50 to 1.72. The covariance between the statistical literacy and the statistical reasoning constructs were allowed to be freely estimated. The estimated covariance was 1.

Table 4.13

Item Parameters for the Correlated Model

	Discrimination				Intercept	
	Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	c	*s.e.*
1*	0.70	0.08	0.00	-----	1.20	0.09
2	1.07	0.11	0.00	-----	1.20	0.11
3*	0.00	-----	0.72	0.12	-0.98	0.12
4	0.00	-----	1.66	0.11	2.04	0.11
5*	0.33	0.14	0.00	-----	3.48	0.15
6*	0.35	0.11	0.00	-----	0.12	0.11
7	0.00	-----	0.99	0.12	0.82	0.12
8	0.00	-----	0.84	0.11	-1.14	0.11
9	1.42	0.12	0.00	-----	1.87	0.11
10	1.24	0.12	0.00	-----	-0.26	0.11
11*	0.00	-----	0.78	0.12	1.41	0.12
12	0.00	-----	0.87	0.12	-0.14	0.11
13	1.09	0.12	0.00	-----	1.04	0.11
14	1.21	0.11	0.00	-----	1.55	0.11
15	0.91	0.11	0.00	-----	-0.42	0.11
16	0.00	-----	1.72	0.11	1.97	0.11
17	0.00	-----	0.89	0.12	-0.45	0.11
18	0.00	-----	1.54	0.11	1.53	0.11
19	1.71	0.11	0.00	-----	1.41	0.11
20	0.00	-----	1.21	0.11	0.44	0.11
21	1.79	0.12	0.00	-----	1.16	0.11
22	0.94	0.11	0.00	-----	1.44	0.11
23	0.00	-----	1.65	0.12	1.48	0.11
24	0.00	-----	0.83	0.12	0.44	0.12
25	1.24	0.12	0.00	-----	0.32	0.10
26	1.13	0.11	0.00	-----	0.06	0.11
27	0.00	-----	0.80	0.12	-0.01	0.11
28*	0.00	-----	0.64	0.12	0.62	0.11
29	0.99	0.11	0.00	-----	-0.40	0.11
30*	0.32	0.11	0.00	-----	0.30	0.11
31	0.00	-----	1.44	0.12	0.10	0.11
32	0.00	-----	1.19	0.12	-0.19	0.11
33*	0.00	-----	0.50	0.11	0.41	0.11
34	1.35	0.11	0.00	-----	0.24	0.10
35	0.00	-----	0.87	0.12	-0.18	0.12
36*	0.44	0.12	0.00	-----	1.06	0.11
37	1.45	0.12	0.00	-----	3.40	0.12
38	1.25	0.12	0.00	-----	0.81	0.11
39	0.00	-----	0.87	0.12	-0.44	0.11

	Discrimination				Intercept	
	Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	*c*	*s.e.*
40	0.00	-----	0.96	0.12	-0.79	0.11

Note. * indicates item with discrimination value lower than 0.8;

Item-level diagnostics statistics are presented in Table 4.14. A total of ten items presented item misfit. Two additional fit indices were computed: the RMSEA was 0.03 and the M_2 (*df* = 739) was 1298.58 (p = .0001).

Table 4.14

Item level diagnostics statistics for the Correlated Model

Item	X^2	*d.f.*	Probability	Item	X^2	*d.f.*	Probability
1	32.31	26	0.1825	21	25.45	23	0.3267
2	**42.63**	**26**	**0.0211**	22	37.11	26	0.0728
3	35.01	27	0.1383	23	24.58	23	0.3743
4	**37.03**	**22**	**0.0234**	24	30.79	28	0.3255
5	5.61	12	0.9345	**25**	**55.76**	**25**	**0.0004**
6	29.86	29	0.4228	**26**	**40.70**	**27**	**0.0439**
7	**43.54**	**26**	**0.0169**	**27**	**44.32**	**27**	**0.0191**
8	38.10	27	0.0760	28	34.17	28	0.1950
9	**35.50**	**23**	**0.0462**	29	25.22	26	0.5077
10	23.96	26	0.5794	30	35.21	31	0.2748
11	29.74	27	0.3251	31	31.37	25	0.1765
12	37.96	27	0.0783	32	22.44	26	0.6652
13	36.86	26	0.0768	33	26.69	29	0.5895
14	26.35	24	0.3348	**34**	**43.48**	**25**	**0.0124**
15	25.89	27	0.5258	35	36.26	26	0.0868
16	15.69	22	0.8316	36	37.05	27	0.0938
17	31.80	27	0.2389	37	17.62	19	0.5493
18	**46.66**	**23**	**0.0025**	**38**	**44.45**	**25**	**0.0096**
19	29.79	23	0.1550	39	29.37	27	0.3423
20	21.81	25	0.6477	40	38.89	27	0.0647

Note. The bold font indicates *p*-value smaller than 0.05.

Figure 4.9 shows the scatterplot of the statistical literacy subscore and statistical reasoning subscore produced by the Correlated Model. The correlation between these subscores was 1.

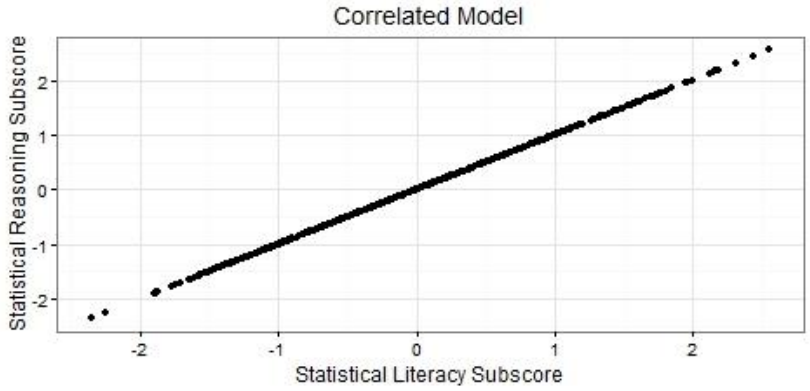


Figure 4.9. Scatterplot of the statistical literacy and statistical reasoning subscores for the Correlated Model.

Cross-loading Model. Table 4.15 provides the item parameters (intercept and discriminations) and standard error of the estimates obtained from fitting the Cross-loading Model to the data. The minimum statistical reasoning item discrimination was 0.25, therefore the cross-loadings from the statistical literacy dimension to the statistical reasoning items was set to be 0.20. In this way, the assumption that each item would have its higher discrimination on its original dimension was met. Item intercept values ranged from –1.14 to 3.47. The estimated item discrimination parameters for the statistical literacy dimension ranged from 0.25 to 1.72. Item discrimination parameters for the statistical reasoning dimension ranged from 0.38 to 1.34. The covariance between the statistical literacy and the statistical reasoning constructs were set to be zero.

Table 4.15

Item Parameters for the Cross-loading Model

	Discrimination				Intercept	
	Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	*c*	*s.e.*
1*	0.62	0.11	0.00	-----	1.19	0.10
2	1.02	0.13	0.00	-----	1.19	0.11
3*	**0.20**	-----	0.53	0.11	-0.97	0.09
4	**0.20**	-----	1.25	0.19	1.86	0.16

	Discrimination				Intercept	
	Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	*c*	*s.e.*
5*	0.26	0.24	0.00	-----	3.47	0.23
6*	0.31	0.09	0.00	-----	0.12	0.08
7	**0.20**	-----	0.86	0.14	0.81	0.10
8*	**0.20**	-----	0.72	0.13	-1.14	0.10
9	1.47	0.18	0.00	-----	1.95	0.17
10	0.99	0.12	0.00	-----	-0.26	0.09
11*	**0.20**	-----	0.71	0.14	1.43	0.11
12*	**0.20**	-----	0.79	0.12	-0.14	0.09
13	1.03	0.13	0.00	-----	1.03	0.11
14	1.07	0.14	0.00	-----	1.51	0.13
15	0.85	0.11	0.00	-----	-0.42	0.09
16	**0.20**	-----	1.34	0.20	1.81	0.16
17*	**0.20**	-----	0.57	0.11	-0.44	0.08
18	**0.20**	-----	1.32	0.19	1.48	0.14
19	1.59	0.18	0.00	-----	1.38	0.14
20	**0.20**	-----	1.07	0.15	0.43	0.10
21	1.72	0.19	0.00	-----	1.15	0.14
22	0.97	0.13	0.00	-----	1.47	0.12
23	**0.20**	-----	1.20	0.17	1.32	0.13
24*	**0.20**	-----	0.57	0.11	0.42	0.08
25	1.30	0.14	0.00	-----	0.33	0.10
26	1.09	0.12	0.00	-----	0.05	0.09
27*	**0.20**	-----	0.63	0.11	-0.02	0.08
28*	**0.20**	-----	0.38	0.11	0.60	0.08
29	0.94	0.11	0.00	-----	-0.41	0.09
30*	0.25	0.09	0.00	-----	0.30	0.08
31	**0.20**	-----	1.23	0.16	0.08	0.10
32	**0.20**	-----	0.96	0.14	-0.19	0.09
33*	**0.20**	-----	0.49	0.11	0.43	0.08
34	1.30	0.14	0.00	-----	0.23	0.10
35	0.20	-----	0.93	0.13	-0.19	0.09
36*	0.47	0.10	0.00	-----	1.07	0.09
37	1.15	0.23	0.00	-----	3.22	0.26
38	1.15	0.13	0.00	-----	0.79	0.10
39*	**0.20**	-----	0.74	0.12	-0.45	0.09
40	**0.20**	-----	0.89	0.13	-0.80	0.10

Note. * indicates item with discrimination value lower than 0.8; The bold font indicates discrimination value (cross-loadings) that were fixed to be 0.20.

Item-level diagnostics statistics are presented in Table 4.16. A total of ten items presented item misfit. The additional fit indices also computed were the RMSEA of 0.04 and the M_2 (*df* = 740) of 1597.72 (p = .0001).

Table 4.16

Item level diagnostics statistics for the Cross-loading Model

Item	X^2	*d.f.*	Probability	Item	X^2	*d.f.*	Probability
1	32.42	28	0.2570	21	29.20	25	0.2551
2	**40.29**	**27**	**0.0479**	22	37.29	26	0.0702
3	34.78	27	0.1442	**23**	**40.57**	**25**	**0.0254**
4	**43.49**	**24**	**0.0087**	24	31.45	29	0.3436
5	7.58	12	0.8176	**25**	**61.57**	**26**	**0.0001**
6	30.78	30	0.4278	**26**	**44.26**	**27**	**0.0194**
7	**43.18**	**27**	**0.0250**	**27**	**44.45**	**28**	**0.0250**
8	36.99	28	0.1188	28	36.53	28	0.1292
9	31.54	24	0.1383	29	27.05	27	0.4625
10	32.73	27	0.2055	30	35.50	31	0.2639
11	30.59	28	0.3345	31	37.09	26	0.0732
12	38.35	28	0.0918	32	25.59	26	0.4873
13	37.62	27	0.0838	33	27.78	29	0.5308
14	27.52	25	0.3294	**34**	**43.46**	**26**	**0.0172**
15	28.20	27	0.4022	35	35.95	26	0.0924
16	23.63	25	0.5420	36	34.00	28	0.2003
17	34.90	27	0.1411	37	19.15	20	0.5137
18	**50.53**	**25**	**0.0018**	**38**	**40.87**	**26**	**0.0319**
19	31.90	25	0.1604	39	30.37	27	0.2972
20	26.43	26	0.4413	40	39.21	27	0.0605

Note. The bold font indicates *p*-value smaller than 0.05.

Figure 4.10 shows the scatterplot of the statistical literacy subscore and statistical reasoning subscore produced by the Cross-loading Model. The correlation between these subscores was 0.640.

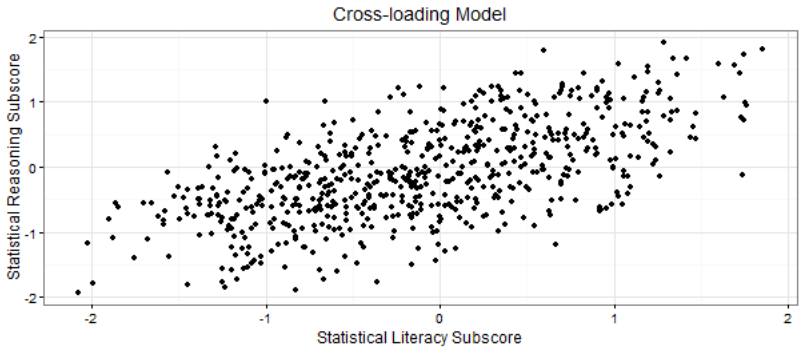


Figure 4.10. Scatterplot of the statistical literacy and statistical reasoning subscores for the Cross-loading Model.

Bi-factor Model A. Table 4.17 provides the item parameters (intercept and discriminations) and standard error of the estimates obtained from fitting the Bi-factor Model A to the data. Item intercept values ranged from –1.22 to 3.70. The discrimination values for the general dimension (statistical knowledge) ranged from 0.34 to 1.79. The estimated item discrimination parameters for the statistical literacy dimension ranged from -0.69 to 1.03. Of all the five IRT models fitted to the data, the Bi-factor Model A was the only model that presented items with negative discrimination. A total of 10 out of the 20 statistical literacy items presented negative item discrimination. Item discrimination parameters for the statistical reasoning dimension ranged from -0.37 to 1.09. This dimension presented 5 items with negative discrimination. The covariance between the statistical literacy and the statistical reasoning constructs were set to be zero. The covariance between the general statistical knowledge construct and the statistical literacy and reasoning constructs were also set to be zero.

Table 4.17

Item Parameters for the Bi-factor Model A

	Discrimination						Intercept	
	General		Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	a_3	*s.e.*	*c*	*s.e.*
1*	0.69	0.11	0.03	0.17	0.00	-----	1.20	0.10

	Discrimination						Intercept	
	General		Literacy		Reasoning			
Item	a_1	*s.e.*	a_2	*s.e.*	a_3	*s.e.*	c	*s.e.*
2*	1.08	0.13	-0.12	0.24	0.00	-----	1.20	0.12
3*	0.68	0.11	0.00	-----	0.52	0.17	-1.01	0.11
4*	1.64	0.19	0.00	-----	0.10	0.23	2.03	0.19
5*	0.35	0.26	0.69	0.47	0.00	-----	3.70	0.37
6*	0.34	0.09	-0.12	0.17	0.00	-----	0.12	0.08
7*	0.99	0.12	0.00	-----	0.50	0.17	0.86	0.11
8*	0.80	0.12	0.00	-----	0.75	0.19	-1.22	0.12
9*	1.43	0.17	0.02	0.27	0.00	-----	1.88	0.16
10*	1.19	0.13	0.06	0.17	0.00	-----	-0.26	0.11
11*	0.77	0.12	0.00	-----	0.15	0.21	1.41	0.12
12*	0.85	0.11	0.00	-----	0.23	0.14	-0.14	0.10
13*	1.12	0.14	-0.20	0.24	0.00	-----	1.06	0.13
14*	1.20	0.15	-0.05	0.18	0.00	-----	1.55	0.14
15	1.06	0.18	1.03	0.29	0.00	-----	-0.49	0.12
16*	1.75	0.21	0.00	-----	-0.23	0.21	2.00	0.18
17*	0.90	0.11	0.00	-----	0.04	0.16	-0.45	0.10
18*	1.52	0.17	0.00	-----	0.06	0.24	1.53	0.18
19*	1.75	0.21	0.22	0.26	0.00	-----	1.43	0.21
20*	1.20	0.13	0.00	-----	0.34	0.18	0.45	0.12
21	1.79	0.21	0.00	0.23	0.00	-----	1.16	0.19
22*	0.95	0.14	-0.33	0.28	0.00	-----	1.47	0.13
23*	1.66	0.20	0.00	-----	-0.11	0.20	1.49	0.20
24*	0.92	0.12	0.00	-----	-0.37	0.15	0.46	0.11
25*	1.29	0.14	0.42	0.30	0.00	-----	0.34	0.11
26*	1.24	0.15	0.68	0.28	0.00	-----	0.07	0.11
27*	0.82	0.11	0.00	-----	-0.09	0.17	-0.02	0.10
28*	0.66	0.11	0.00	-----	-0.14	0.15	0.62	0.10
29*	1.05	0.13	0.66	0.21	0.00	-----	-0.43	0.10
30*	0.34	0.09	-0.36	0.17	0.00	-----	0.31	0.09
31*	1.42	0.14	0.00	-----	0.42	0.16	0.11	0.14
32*	1.18	0.13	0.00	-----	0.11	0.15	-0.19	0.12
33*	0.51	0.12	0.00	-----	1.09	0.29	0.51	0.12
34*	1.45	0.15	-0.37	0.21	0.00	-----	0.25	0.14
35*	0.90	0.12	0.00	-----	0.89	0.24	-0.19	0.12
36*	0.48	0.11	-0.26	0.18	0.00	-----	1.09	0.10
37*	1.50	0.29	-0.69	0.35	0.00	-----	3.62	0.37
38*	1.35	0.16	-0.53	0.26	0.00	-----	0.86	0.12
39*	0.84	0.11	0.00	-----	0.52	0.16	-0.45	0.11
40*	0.94	0.13	0.00	-----	0.53	0.18	-0.81	0.13

Note. * indicates item with discrimination value lower than 0.8.

Item-level diagnostics statistics are presented in Table 4.18. A total of ten items presented item misfit. Two additional fit indices were computed: the RMSEA was 0.03 and the M_2 (df = 700) was 1093.55 (p = .0001).

Table 4.18

Item level diagnostics statistics for the Bi-factor Model A

Item	X^2	*d.f.*	Probability	Item	X^2	*d.f.*	Probability
1	32.05	25	0.1561	21	25.56	22	0.2705
2	**42.14**	**25**	**0.0173**	22	37.41	25	0.0526
3	34.42	26	0.1245	23	24.33	22	0.3291
4	**36.87**	**21**	**0.0174**	24	30.76	27	0.2804
5	5.62	11	0.8979	**25**	**57.80**	**25**	**0.0002**
6	29.95	28	0.3673	**26**	**39.89**	**26**	**0.0399**
7	**44.19**	**25**	**0.0103**	**27**	**42.75**	**26**	**0.0205**
8	36.61	25	0.0626	28	34.21	27	0.1597
9	35.14	23	0.0503	29	24.17	25	0.5106
10	24.03	25	0.5187	30	34.82	30	0.2485
11	29.75	26	0.2771	31	30.83	24	0.1582
12	37.80	26	0.0631	32	22.56	25	0.6040
13	**37.66**	**25**	**0.0498**	33	26.15	28	0.5658
14	26.30	23	0.2862	**34**	**42.18**	**24**	**0.0123**
15	25.44	26	0.4955	35	34.19	25	0.1037
16	15.14	21	0.8166	36	36.61	26	0.0808
17	31.73	26	0.2017	37	17.76	18	0.4728
18	**46.72**	**22**	**0.0016**	**38**	**44.12**	**24**	**0.0074**
19	30.03	22	0.1173	39	28.04	26	0.3556
20	21.74	24	0.5959	40	38.00	26	0.0604

Note. The bold font indicates *p*-value smaller than 0.05.

Figure 4.11 shows the scatterplot of the statistical literacy subscore and statistical reasoning subscore. These subscores have had a very low correlation of 0.078. From all the five IRT models, this is the only model in which the subscores appear to be almost uncorrelated.

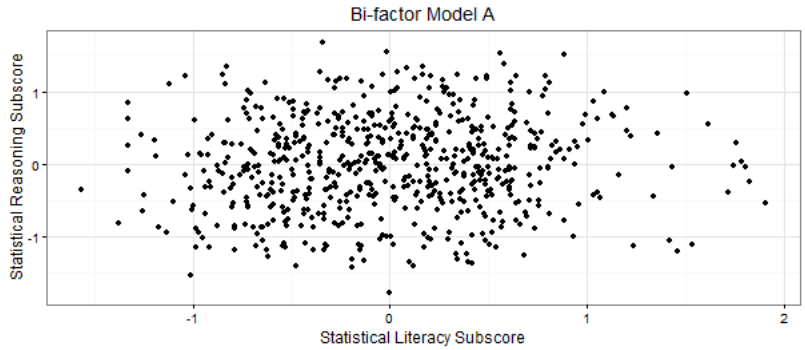


Figure 4.11. Scatterplot of the statistical literacy and statistical reasoning subscores.

Figure 4.12 shows the scatterplots between the estimated general dimension subscore and the estimated statistical literacy and statistical reasoning subscores. The correlation between the subscores on the observed general dimension and the statistical literacy and statistical reasoning subscores were 0.033 and 0.097, respectively.

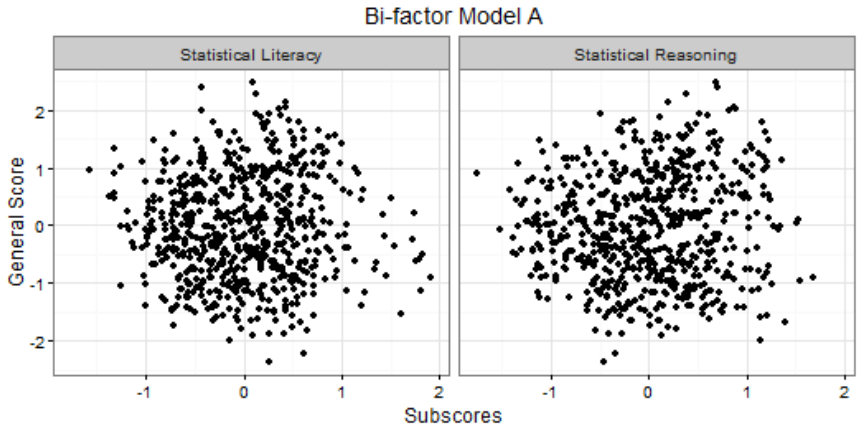


Figure 4.12. Scatterplots of the score from the general dimension and the statistical literacy and statistical reasoning subscores.

4.4.3 Comparisons of the Five Fitted IRT Models. The five IRT models (Unidimensional, Uncorrelated, Correlated, Cross-loadings, and Bi-factor A) were compared using item- and model-level measures of fit to find the best model(s) that represents the learning goals of statistical literacy and the construct of statistical reasoning.

Item-level measures. Examination of the S-X^2 statistic indicated misfit of ten items for the Unidimensional Model, Correlated Model, Cross-loading Model, and Bi-factor Model A. The Uncorrelated Model presented misfit for nine items, according to the S-X^2 statistic. A couple of different items were flagged by each model, but nine items (items 2, 4, 7, 13, 18, 25, 26, 27, 34, and 38) were consistently flag by all IRT models (see Table 4.19).

Table 4.19

Items that presented misfit according to the S-X^2 statistic for each model.

Item	Unidimensional	Uncorrelated	Correlated	Cross-loading	Bi-factor A
2	X	X	X	X	X
4	X	X	X	X	X
7	X	X	X	X	X
9	X		X		
13	X	X	X	X	X
18	X	X	X	X	X
23				X	
25	X	X	X	X	X
26	X	X	X	X	X
27	X	X	X	X	X
34	X	X	X	X	X
38	X	X	X	X	X

Model-level measures. Model-based measures of fit were also computed for each of the five IRT models (see Table 4.20): -2lnL, AIC, BIC, RMSEA, and the *p*-value for the M_2 statistic.

Table 4.20

Model-based Measures of Fit

Fit Measures	Unidimensional	Uncorrelated	Correlated	Cross-loading	Bi-factor A
p-value for M_2	0.0001	0.0001	0.0001	0.0001	0.0001
RMSEA	0.03	0.04	0.03	0.04	0.03
AIC	29648.41	30294.51	29650.42	30143.69	29578.91
BIC	30009.12	30655.21	30015.63	30504.39	30119.96
-2lnL	29488.41	30134.51	29488.42	29983.69	29338.91

Table 4.20 shows the estimate of the M_2 fit statistic which tests if the model fits the data considering all items together. The results were significant ($p < 0.001$) for all five models. Based on these results, the M_2 statistic indicated significant misfit to the data for all models. The RMSEA values indicated close fit to the data (values smaller than 0.06) for all five IRT models based on guidelines suggested by Browne and Cudeck (1993).

Both AIC and BIC rank ordered the IRT models very similarly. The AIC statistic ranked the models according to the first column of Table 4.21. The BIC statistic ranked the models according to the second column of Table 4.21.

Table 4.21

Rank order of Models based on the AIC and BIC statistics

AIC	BIC
Bi-factor A	Unidimensional
Unidimensional	Correlated
Correlated	Bi-factor A
Cross-loading	Cross-loading
Uncorrelated	Uncorrelated

The Unidimensional Model, Correlated Model, and Bi-factor Model A were the three models with the smallest AIC and BIC values. However, AIC and BIC ranked these three models differently (see Table 4.21). On the other hand, the Cross-loading Model and Uncorrelated Model were ranked in the same way by AIC and BIC (see Table 4.21).

Nested models can also be compared using a likelihood ratio test. There are two groups of nested models: models with an overall dimension (Unidimensional and Bi-factor A) and bi-dimensional models (Uncorrelated, Correlated and Cross-loading). Table 4.22 shows the likelihood ratio test for each comparison within each group.

Table 4.22

Likelihood ratio tests

Reduced model	Full model	ΔG^2	p-value	Preferred model
Models with an overall dimension				
Unidimensional	Bi-factor A	149.50	< 0.001	Bi-factor A
Bi-dimensional models				
Uncorrelated	Correlated	646.09	< 0.001	Correlated
Cross-loading	Correlated	495.27	< 0.001	Correlated

The results of the likelihood ratio tests indicated that the Bi-factor Model A likely fits the data better than the Unidimensional. In addition, the results also supported the assumption that the Correlated Model likely fits the data better than the Uncorrelated Model and Cross-loading Model.

4.4.4 Subscore Analysis. The subscores from the IRT models were evaluated in regard to distinctiveness and reliability. The Haberman analysis (Haberman, 2008) was also used to evaluate if reporting the statistical literacy and statistical reasoning subscores provided distinct information over and above that given by the total score

Subscore Correlation. The second column in Table 4.23 shows the correlation between the statistical literacy and statistical reasoning subscores for each of five models: Uncorrelated Model, Correlated Model, Cross-loading Model, and Bi-factor Model A. The Bi-factor Model A was the only model that produced a very small correlation between the two subscores (r = 0.078). The Cross-loading Model and Uncorrelated Model produced a modest correlation of 0.64 and 0.68, respectively. The Correlated Model produced a perfect subscore correlation of 1. In other words, the Correlated Model produced, for each student, the same statistical literacy and statistical reasoning subscores.

Table 4.23

Correlation within subscores and between subscores and the total score from the Unidimensional Model.

	Correlation		
Model	Literacy & Reasoning	Total Score & Literacy	Total Score & Reasoning
Uncorrelated	0.684	0.902	0.915
Correlated	1	0.999	0.999
Cross-loading	0.640	0.924	0.870
Bi-factor A	0.078	0.051	0.170

To evaluate the distinctiveness and reliability of the statistical literacy and statistical reasoning subscores, these subscores were compared to the total score from the Unidimensional Model. The third and fourth columns on Table 4.23 presents the correlations of the total score from the Unidimensional Model with the statistical literacy subscore and statistical reasoning subscore from each of the five models. The Bi-factor Model A again is the only model whose subscores were not correlated with the total unidimensional score. All the other models, on the other hand, present high to very high correlations with the total unidimensional score.

Subscore Reliability. Table 4.24 shows the reliability of each subscore for each model and the reliability of the difference between subscores for each model. The reliability of the IRT subscores was computed using Equation 3.8. The last column of Table 4.24 presents again the correlation between the statistical literacy and statistical reasoning subscores. This last column is the same as the first column in Table 4.23, but it was added here to facilitate interpretation.

Table 4.24

Reliability of each subscore, of the difference within subscores and correlation between subscores.

	Reliability			Correlation
Model	Literacy	Reasoning	Difference between subscores	Literacy & Reasoning
Uncorrelated	0.736	0.763	0.207	0.684
Correlated	0.878	0.878	-	0.999
Cross-loading	0.752	0.696	0.235	0.640
Bi-factor A	0.356	0.408	0.329	0.078

The Correlated Model produced the highest values of subscore reliability followed by the Uncorrelated Model. The reliability of the subscores were fairly similar within each model. The difference in reliability between the statistical literacy and statistical reasoning subscores was smaller than 0.06 for all models. The Cross-loading Model was the only model that produced a slightly lower reliability for the statistical reasoning subscore when compared to the statistical literacy subscore for this same model. The model that presented the smallest subscore reliability was the Bi-factor Model A.

The reliability for the Unidimensional Model was 0.879 and the reliability for the general subscore for the Bi-factor Model A was 0.871.

Table 4.24 also shows the reliability of the difference between subscores. The reliability of the difference did not produce a meaningful result for the Correlated because the subscores for statistical literacy and statistical reasoning were the same. As the correlation between the subscores increases (see last column in Table 4.24), the reliability of the difference between subscores decreases. Therefore, the Bi-factor Model A presented the highest reliability of the difference (0.331) because this model produced the smallest correlation between the statistical literacy and statistical reasoning subscores (r = 0.078). As can be seen from Table 4.24 all models produced very low reliability of the difference between subscores.

Haberman Analysis. The Haberman Analysis verified for each model if the statistical literacy subscore and statistical reasoning subscore contained more information than the total score from the Unidimensional Model.

The correlation of each of the subscores with the total score from the Unidimensional Model is reported on Table 4.25. It can be seen from this table that both subscores for the Uncorrelated, Correlated and Cross-loading models were highly correlated with the unidimensional total score. The same is true for the correlations between the unidimensional total score with the

general factor from the bi-factor model. The correlation of each subscore with the unidimensional total score was smaller only for the Bi-factor Model A.

Table 4.25

Correlation of each subscore to the total score from the Unidimensional Model.

Model	Literacy	Reasoning
Uncorrelated	0.9023	0.9150
Correlated	0.9999	0.9999
Cross-loading	0.9240	0.8696
Bi-factor A	0.0506	0.1699
Bi-factor A - General	0.9969	

Table 4.26 shows the results of the Haberman Analysis. This table lists the proportional reduction in mean square error (PRMSE) for each of the five IRT models. For a subscore to have added value over the total score, the PRMSE(S_x) should be greater than the PRMSE(S_z). The IRT models that produced the most difference between PRMSE(S_x) and PRMSE(S_z) were Bi-factor Model A and the Cross-loading Model. The Uncorrelated Model also produced an increase in PRMSE(S_x) but not as much as the Bi-factor Model A and Cross-loading Model.

Table 4.26

PRMSEs for each subscore for of the five IRT models

Model	Learning Goal	PRMSE(S_z) Total Score	PRMSE(S_x) Subscore	Difference
Uncorrelated	Literacy	0.6774	0.7356	0.0583
	Reasoning	0.7068	0.7631	0.0563
Correlated	Literacy	0.8785	0.8939	0.0098
	Reasoning	0.8785	0.8939	0.0098
Cross-loading	Literacy	0.6904	0.7523	0.0619
	Reasoning	0.5903	0.6959	0.1056
Bi-factor A	Literacy	0.3129	0.3555	0.0427
	Reasoning	0.1064	0.4076	0.3013

4.5 Chapter summary

This chapter presented the results from the expert review and think-aloud interviews performed during the development process of the REALI assessment. The chapter also described the results from the pilot, field test and the data analysis to answer the three research questions posed by this study. The next chapter presents the discussion of the results of the study.

Chapter 5

Discussion

This chapter provides a summary of the study followed by the synthesis and discussion of the results. This includes a critique of the expert reviews of the initial set of items, as well as a discussion of what was learned about the psychometric properties of the REALI assessment. In addition, this chapter provides responses to each of the research questions posed in this study. The chapter concludes with limitations and implications for teaching and future research.

5.1 Summary of the study

The purpose of this study was to distinguish between and explore the relationship between statistical literacy and statistical reasoning. Three research questions were posed to structure the research: (1) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony? (2) What measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction? (3) What measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning?

The REALI assessment was developed to concurrently measure statistical literacy and statistical reasoning. This instrument is comprised of 40 items, with 20 items measuring statistical literacy and 20 item measuring statistical reasoning. Eight areas of learning are assessed by this instrument: (1) representations of data, (2) measures of center, (3) measures of variability, (4) study design, (5) hypothesis testing and *p*-values, (6) confidence intervals, (7) bivariate data, and (8) probability.

As suggested by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), careful attention was given to how the scores from the REALI instrument would

be interpreted. To support the intended inferences and uses of the scores different types of validity evidence were gathered throughout the development process: expert reviews, response process interviews with students, a pilot test, a field test, and a psychometric analysis.

Standard 1.2 from the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) suggests that the constructs being measured by the assessment should be clearly described. Therefore, the first step in the development process was establishing the working definitions of statistical literacy and statistical reasoning. The working definition of statistical literacy was based on the definition from Ziegler (2014) and statistical reasoning was defined based on the definitions from Garfield and Ben-Zvi (2008) and delMas (2002, 2004):

- Statistical literacy items assess students' ability to recall a definition, describe or interpret basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will not require that students make connections between them (recall information will be sufficient).
- Statistical reasoning items assess students' ability to make connections among statistical concepts, create mental representations of statistical problems, and explain relationships between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, statistical reasoning items require higher order thinking and higher cognitive load than statistical literacy items.

The development process continued with the elaboration of a test blueprint. The basis of the REALI assessment was the items from the GOALS instrument, which measures statistical reasoning, and items from BLIS, which measures statistical literacy. Items were grouped into areas of learning based on the content addressed by each item. At the end, eight areas of learning were

identified (representations of data, measures of center, measures of variability, study design, hypothesis testing and *p*-values, confidence intervals, bivariate data, and probability).

Further investigations were done to verify if the items aligned with the working definitions of statistical literacy and statistical reasoning. The behaviors needed to answer each item correctly were identified and used to classify each item as being a statistical literacy or a statistical reasoning item. A categorization of items into these two categories was also done by experts in the field of statistics education, this was part of the expert review. Experts were also invited to critique the items. As specified by Standard 3.5 (AERA, APA, & NCME, 1999), feedback from the experts was used as validity evidence to assure quality of items and item's categorization. Experts' feedback was also used to make changes to the instrument.

Additional validity evidence was gathered from the think-aloud interviews with students. These interviews were conducted to better understand how students responded to each of the items. Their responses were used to verify if items were behaving as intended and to clarify the categorization of some items. Changes were made to the items based on analysis of students' responses.

The pilot test was the next step in the development process. A total of 237 students completed the REALI assessment as part of the pilot study. Item characteristics such as item difficulty and item discrimination were estimated to verify how the items were performing. Additional changes were made to the instrument based on the pilot testing. The resulting instrument was the final version of the assessment. This version was used in the field test with 671 students from 16 universities and colleges around the United States and Canada. Students' answers from the field test were used to explore the relationship between statistical literacy and statistical reasoning and to answer the three research questions posed in this study.

5.2 Discussion of the Results

5.2.1 Expert Review – Categorization of Items. During the expert review process, the experts were asked to categorize each item into two groups: group 1 (statistical literacy) and group 2 (statistical reasoning). In general, for most of the items, the experts' classifications were the same as the classification of the researcher. Only nine out of the 52 items had poor level of agreement – an item presents poor level of agreement when less than half of the experts agree with the categorization of the researcher for that item.

Overall, there was more agreement between the experts and the researcher for the statistical literacy items than for the statistical reasoning items. Twenty-seven out of the 28 statistical literacy items, or 96% of the statistical literacy items, produced moderate to high level of agreement between the experts and the researcher – an item produced a moderate to high level of agreement when half or more of the experts agree with the categorization of the researcher for that item. On the other hand, only 16 out of the 24 statistical reasoning items, or 66% of the statistical reasoning items, had a moderate to high level of agreement. These results suggest that it was easier for the experts to agree on the categorization of statistical literacy items than to agree on the categorization of statistical reasoning items. A possible reason why this happened could be related to the definitions of statistical literacy and statistical reasoning items used in this study. For instance, to categorize an item as a statistical reasoning item, the reviewers had to carefully examine each item to verify how many statistical concepts were being addressed and then examine if these concepts needed to be connected to answer the question correctly. Therefore, recognizing statistical reasoning items demanded more steps than recognizing a statistical literacy item. Two additional problems related to statistical reasoning items were that (1) the definition of a "statistical concept" might not have been clear to some of the reviewers and (2) for some items the relationship between

concepts happened in the alternative options and not on the stem of the problem. These issues will be explored in the following paragraphs.

A total of nine items out of the 52 items showed little agreement between the researcher and the experts. Eight of these items were categorized as statistical reasoning items by the researcher, but the experts disagreed with this categorization. All the items that had 0% agreement between the researcher and the experts were items that presented the statistical reasoning part in the alternative options and not in the stem of the item. For instance, items 3A, 5D, 7D, 8F, and 4G were categorized as statistical reasoning items by the researcher because the alternative options of these items addressed more than one statistical concept. Therefore, when students go through the alternative options to answer the item, they are forced to make connections between more than one statistical concept thus exhibiting statistical reasoning.

None of the expert reviewers classified the items mentioned above (items 3A, 5D, 7D, 8F, and 4G) as statistical reasoning items. A possible reason why the researchers disagreed with the researcher on the categorization of these items could be that the experts did not consider that statistical reasoning could happen while students were reading the alternative options. It is possible that the researchers focused only on the stem of the items.

Items 7F and 4E were the other two items that presented a low agreement level between the researcher and the experts. Item 7F was designed to assess if the students know how increasing the sample size affects the p-value, all else being equal. Item 4E assessed if students were able to reason about how the width of a confidence interval is related to sample size. Therefore, these items were categorized as statistical reasoning items by the researcher because students make connections between more than one statistical concept. For item 7F, students have to make connections between the sample size and the p-value and, for Item 4E, connections between sample size and confidence intervals.

An explanation as to why most of the researchers classified Item 7F and 4E as statistical literacy items instead of a statistical reasoning items could be related to how they interpret the term "statistical concepts". For instance, the working definition of statistical reasoning used in the categorization of items stated that statistical reasoning items needed to address more than one *statistical concept* and require making connections between them. So a possible reason why most of the researchers disagreed with the categorization of items 7F and 4E could be that they did not consider "sample size" as a statistical concept.

5.2.2 Research Questions

Three research questions were posed in this study. The first research question to be answered was *what measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony?* The second research question was *what measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction?* And the final research question was *what measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning?*

Research Question 1. To answer the first research question, five IRT models were fitted to students' responses: (1) Unidimensional Model, (2) Uncorrelated Model, (3) Correlated Model, (4) Cross-loading Model, and (5) Bi-factor Model A.

The first model was a unidimensional model with only one overall dimension (statistical knowledge). The next three models (Uncorrelated Model, Correlated Model, and Cross-loading Model) were bi-dimensional models, each composed of two dimensions: a statistical literacy dimension and a statistical reasoning dimension. Finally, the last model (Bi-factor Model A) was a

bi-factor model composed of three dimensions: a general dimension of statistical knowledge and two sub-dimensions of statistical literacy and statistical reasoning.

These models were compared at both the item- and model level. At the item-level the $S\text{-}X^2$ statistics was used and at the model-level the M_2, RMSEA, AIC, BIC, and Likelihood Ratio Test (LRT) were used. Table 5.1 shows the models that each of the statistics favored. Favored models are shown with a checkmark symbol (✓).

Table 5.1

Summary of model comparison

					Likelihood Ratio Test	
Model	$S\text{-}X^2$	M_2	RMSEA	AIC & BIC	Group 1*	Group 2♦
Unidimensional*	✓		✓	✓	☐	
Uncorrelated♦	✓		✓			☐
Correlated♦	✓		✓	✓		☑
Cross-loading♦	✓		✓			☐
Bi-factor A*	✓		✓	✓	☑	

Note. * indicates models in group 1 and ♦ indicates models in group 2.

The $S\text{-}X^2$ statistic flagged about the same number of items for all models, thus there does not appear to be a better model based on this item-level statistic. The M_2 statistic indicated significant misfit to the data for all models. However, like all chi-square statistics, the M_2 is overly sensitive to small deviations from the model. So even though M_2 suggested a statistically significant deviation from the model, the RMSEA values indicated that this deviation wass not of practical significance. The RMSEA also did not favor any specific model since all five models produced evidence of close fit.

The other fit measures used to compare the models were AIC, BIC, and the LRT. The AIC and BIC fit statistics ranked the models in a similar way with the initial three models (models with the smallest AIC and BIC) being the Unidimensional, Correlated Model, and Bi-factor Model A. The LRT was used to compare nested models. The first group of models compared were the

Unidimensional and Bi-factor Model A (group 1). The Bi-factor Model A model was preferred over the Unidimensional Model. The second group of models compared were the Uncorrelated, Correlated, and Cross-loading models (group 2). The LRT favored the Correlated Model over the Uncorrelated and Cross-loading models.

Taking into consideration the discussion of model fit presented above, three models were preferred: the Unidimensional Model, the Correlated Model and the Bi-factor Model A. However, the Bi-factor Model A estimated 40 more parameters than the Unidimensional Model and 39 more parameters than the Correlated Model. Therefore, taking into consideration parsimony the Unidimensional Model and the Correlated Model seem to be the best models to represent the construct of statistical literacy and the construct of statistical reasoning given the criteria of fit and parsimony.

Research Question 2. The second research question posed in this study was *what measurement model best represents the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction?* To answer this question, the subscores of the IRT models were evaluated in terms of reliability and distinctiveness. The REALI instrument was designed to measure statistical knowledge, specifically two statistical learning goals: statistical literacy and statistical reasoning. Based on the answer to the first research question, it is clear that all the models presented evidence of close fit, therefore, there is some evidence that the statistical literacy and statistical reasoning learning goals might not form a unidimensional construct. Therefore, the REALI instrument can be used to obtain, for each person, two subscores: a statistical literacy subscore and a statistical reasoning subscore. However, Standard 1.12 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) states that evidence of distinctiveness should be reported if a test provides more than one score. In addition, standards 2.1 and 5.2 request that evidence of reliability, validity and comparability be reported not

only for the total score, but also for subscores. Therefore, substantial evidence should be established to support the use and interpretation of subscores.

Each of the IRT models estimated person-scores (or subscores or ability scores) for each student. The Unidimensional Model produced one score for each student representing each student's statistical knowledge. The Uncorrelated Model, Correlated Model, and Cross-loading Model did not produce one score for each student. Instead, these three models produced two subscores for each student: a statistical literacy subscore and a statistical reasoning subscore. Finally, Bi-factor Model A produced three subscores for each student: a general subscore representing statistical knowledge and two additional subscores representing statistical literacy and statistical reasoning. Table 5.2 shows the correlation between all scores and subscores from all models.

It was expected that the statistical literacy subscore and statistical reasoning subscore would be similar across models because each subscore mainly consisted of the same items. Items 1, 2, 5, 6, 9, 10, 13, 14, 15, 19, 21, 22, 25, 26, 29, 30, 34, 36, 37, and 38 for the statistical literacy subscores and items 3, 4, 7, 8, 11, 12, 16, 17, 18, 20, 23, 24, 27, 28, 31, 32, 33, 35, 39, and 40 for the statistical reasoning subscore. However, an unexpected result was that some of the bi-dimensional and the bi-factor model produced a very high correlation between their subscores and the score from the Unidimensional Model. Going forward, the score from the Unidimensional Model will be referred to as the total unidimensional score.

Table 5.2

Table of correlations between scores from each of the five IRT models.

Score	1	2	3	4	5	6	7	8	9	10
1. Unidimensional-General	—	0.902	0.915	1.000	1.000	0.924	0.870	0.997	0.051	0.170
2. Uncorrelated-Literacy	0.9023	—	0.684	0.911	0.893	0.997	0.610	0.908	0.091	0.043
3. Uncorrelated-Reasoning	0.9150	0.684	—	0.907	0.923	0.713	0.994	0.898	-0.009	0.335
4. Correlated-Literacy	0.9999	0.902	0.915	—	1	0.934	0.869	0.997	0.051	0.170
5. Correlated-Reasoning	0.9999	0.902	0.915	1	—	0.934	0.869	0.997	0.051	0.170
6. Cross-loading-Literacy	0.9240	0.997	0.713	0.932	0.916	—	0.640	0.929	0.087	0.057
7. Cross-loading-Reasoning	0.8696	0.610	0.994	0.860	0.879	0.640	—	0.850	-0.019	0.362
8. Bi-factor A-General	0.9969	0.908	0.898	0.997	0.996	0.929	0.850	—	0.033	0.097
9. Bi-factor A-Literacy	0.0506	0.091	-0.009	0.053	0.049	0.087	-0.019	0.033	—	0.078
10. Bi-factor A-Reasoning	0.1699	0.043	0.335	0.160	0.181	0.057	0.362	0.097	0.078	—

The models which presented very similar scores when compared to the total unidimensional score were the Correlated Model and the Bi-factor Model A. For the Correlated Model, the correlation between the statistical literacy subscore with the total unidimensional score was 0.999. The correlation between the statistical reasoning subscore and the total unidimensional score was also 0.999. In addition, the correlation between the standard error of both subscores were also very highly correlated with the standard error of the total unidimensional score—correlation of 0.999. Therefore, it seems that the Correlated Model shares too much information with the Unidimensional Model leading to subscores and standard errors that are almost the same. In addition, the Correlated Model produced for each person the same value for the statistical literacy subscore and the statistical reasoning subscore. Thus the subscores from the Correlated Model do not help to differentiate between students' statistical literacy and statistical reasoning ability. For these reasons, the Correlated Model will not be included in further discussion regarding subscores analysis.

The Bi-factor Model A also produced a very high correlation between its general statistical knowledge subscore and the total unidimensional score from the Unidimensional Model—correlation of 0.997 (see Figure 5.1). In addition, the standard errors between these scores were also highly correlated (0.993). Therefore, it seems that the total score from the Unidimensional Model and the general statistical knowledge score from the Bi-factor Model A are practically the same. This high correlation supports the assumption that the general statistical knowledge dimension, from the Bi-factor Model A, explains most of the variance in students' responses. This high correlation also explains why the statistical literacy subscore and the statistical reasoning subscore, from the Bi-factor Model A, had a very low correlation with the general factor from the Bi-factor Model A (see Figure 4.21) and with the total unidimensional score (see Figure 5.2). In addition, this was the only IRT model that produced negative item discrimination values. Taking

into consideration the discussion above, it seems that the Bi-factor Model A shares too much information with the Unidimensional Model leading to subscores and standard errors that are almost the same. Thus the Bi-factor Model A will not be included in further discussion regarding subscores analysis.

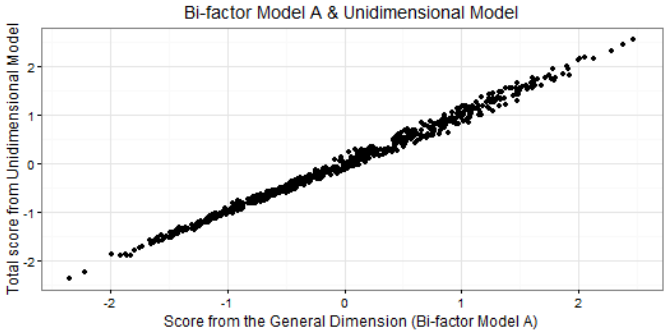


Figure 5.1. Scatterplot of the scores from the general dimension from the Bi-factor Model A and the total score from the Unidimensional Model.

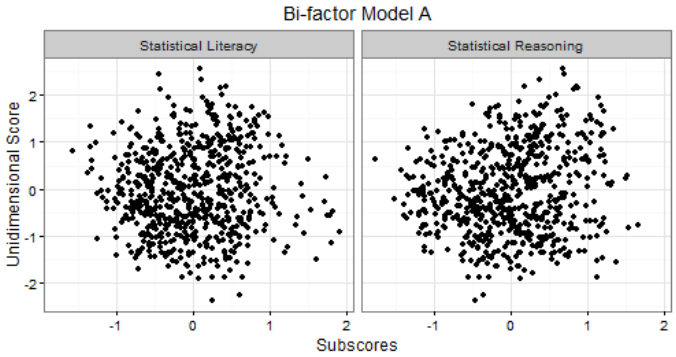


Figure 5.2. Scatterplot of the scores from the general dimension from the Bi-factor Model A and the total unidimensional score from the Unidimensional Model.

Based on the discussion above, the Correlated Model and the Bi-factor Model A produced almost the same results as the Unidimensional model. Therefore, only the Unidimensional Model, Uncorrelated Model, and Cross-loading Model will be considered to answer the second research question.

To evaluate how reliable the statistical literacy and statistical reasoning subscores are, the reliability of each subscore was computed for all models (see Table 4.27). The models that provided

more reliable subscores were the Uncorrelated Model (statistical literacy–0.736 and statistical reasoning–0.763) and Cross-Loading Model (statistical literacy–0.752 and statistical reasoning–0.696). Therefore, students' statistical literacy and statistical reasoning subscores obtained from these models presented moderate to high reliability.

To evaluate the distinctiveness of the statistical literacy and statistical reasoning subscores, the correlation between subscores (see Table 4.26) and the reliability of the difference between subscores was computed for each model (see Table 4.27). There is an inverse relationship between these two measures. As the correlation between the subscores increases, it becomes more difficult to differentiate between the statistical literacy subscore and the statistical reasoning subscore leading to a decrease in the reliability of the difference between subscores. The model that presented a higher level of distinctiveness between the statistical literacy and statistical reasoning subscore was the Cross-loading Model followed by the Uncorrelated Model.

The Haberman analysis (Haberman, 2008) was used to evaluate if reporting two subscores, one for statistical literacy and one for statistical reasoning, had added value over reporting only the total unidimensional score from the Unidimensional Model. Among other things, this analysis takes into account the reliability of each subscore and how correlated the subscores are with each other and with the total unidimensional score. The Cross-loading Model produced the most evidence supporting the use of subscores over the total unidimensional score. These results are aligned with the correlations between each of the subscores and the total unidimensional score. The subscores that produced smaller correlation with the total unidimensional score were the subscores with highest difference in PRMSE and thus more evidence of distinction from the total unidimensional score.

Based on the results and discussion above, it seems that evidence was found supporting the distinctiveness and reliability of subscores. The Cross-loading Model presented the highest

evidence of distinction and also presented evidence that the statistical literacy and statistical reasoning scores can be measure reliably. In addition, the Cross-loading Model was the model which presented the highest evidence that the subscores provide information that is over and above the information provided by the total unidimensional score. It seems therefore that the Cross-loading model is the best for representing the construct of statistical literacy and the construct of statistical reasoning given the criteria of reliability and distinction.

Research Question 3. The third research question posed in this study was *what measurement model is most useful for understanding the constructs of statistical literacy and statistical reasoning?*

The Cross-loading Model presented evidence of close fit do the data according to the S-X^2 and RMSEA statistics. This model also estimated the same number of parameters as the Unidimensional Model. Therefore, the Cross-loading Model was one of the most parsimonious models. Finally, the Cross-loading Model presented most evidence supporting the distinctiveness and reliability of the statistical literacy and statistical reasoning subscores. Thus it seems that the Cross-loading Model is the model that is most useful for understanding how statistical literacy and statistical reasoning are related.

The Cross-loading Model used direct effects from the statistical literacy dimension to the statistical reasoning items (Figure 5.3). In addition, this model assumed that these direct effects would be the same for all items and would be smaller than all the effects of the statistical reasoning dimension on the statistical reasoning items. In other words, the statistical reasoning dimension would have the highest effect on the statistical reasoning items.

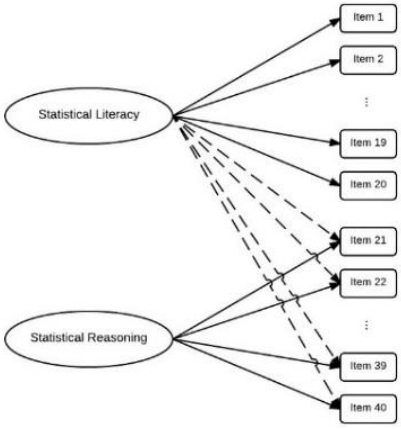


Figure 5.3. Bi-dimensional IRT models.

The cross-loadings and assumptions from the Cross-loading model support the theory from Garfield and Ben-Zvi (2008) of a hierarchy between statistical literacy and statistical reasoning, with statistical literacy being the basis for statistical reasoning. The results found in this study also supported the second theoretical model proposed by delMas (2002). In his model, statistical literacy overlaps entirely with statistical reasoning; however, there is an independent part of statistical literacy (see Figure 5.4).

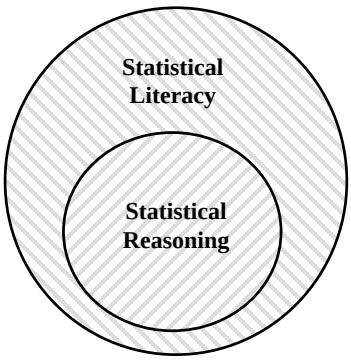
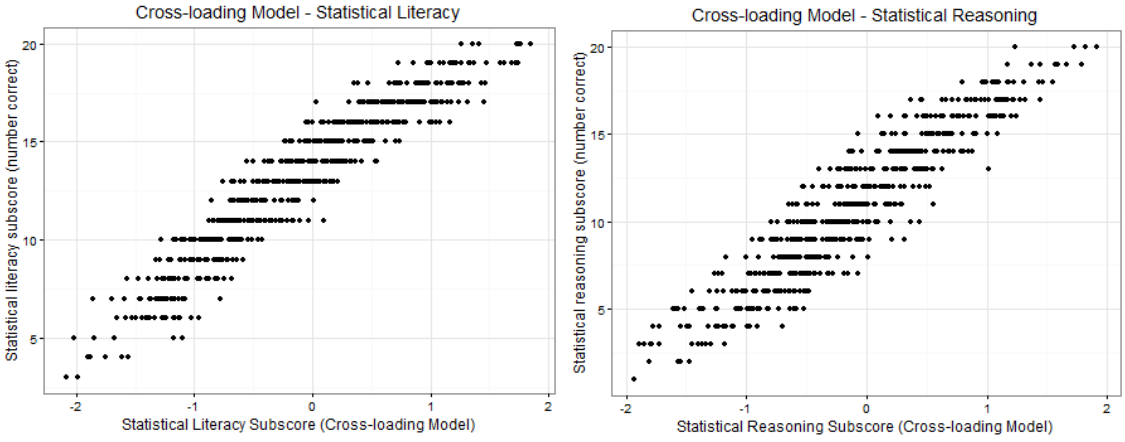


Figure 5.4. Part of the model representing the relationship among statistical literacy and statistical reasoning proposed by delMas (2002)

Based on the information above, the Cross-loading Model provides subscores that are useful to explain the relationship between statistical literacy and statistical reasoning. However, the usefulness of subscores can only be applied in the IRT subscore and not for raw scores using the

number of items correct. The next subsection reports information about how the statistical literacy and statistical reasoning subscores relate to the raw subscores when the number of items correct is used.

Relationship between the Cross-loading Model and the raw scores. The statistical literacy subscore from the Cross-loading Model produced a high correlation (0.950) with the raw statistical literacy subscore using the number of items correct (see Figure 5.5a). Similarly, the statistical reasoning subscore also was also very highly correlated (0.934) with the raw statistical reasoning subscore using the number of items correct (see Figure 5.5b).



(a) (b)

Figure 5.5. Relationship between IRT subscores from the Cross-loading Model and the raw scores.

From both plots in Figure 5.5, it can be seen that the relationship between the raw subscore and the IRT subscore from the Cross-loading Model is not one to one. For instance, a couple of students received a raw literacy subscore of 10, however, each of these students receive a different statistical literacy subscore under the Cross-loading IRT Model. As can be seen in Figure 5.5a the IRT subscores equivalent to a raw literacy subscore of 10 was anywhere from -1.3 to -0.5.

In terms of distinctiveness and reliability, the raw subscores performed differently than the subscores from the IRT model. The raw subscores take into account only the number of items each

student correctly answered; however, the IRT subscores take into account which of the items the student correctly answered, the discrimination, and difficulty of such items. The relationship between the subscores also changes under the item response theory framework. The correlation between the raw statistical literacy and raw statistical reasoning subscores was 0.764. On the other hand, the correlation between the subscores from the Cross-loading model decreased to 0.640. The reliability of the difference between subscores is higher for the Cross-loading subscores (0.247) than for the raw subscores (0.047). Therefore, the subscores from the IRT model are more distinct than the raw subscores.

On the other hand, the reliability of the raw scores was slightly higher than the reliability of the subscores form the Cross-loading model. For the raw subscores, the reliability for the statistical literacy subscore and statistical reasoning subscores were 0.762 and 0.782, respectively. For the Cross-loading Model, the reliabilities for the statistical literacy and statistical reasoning subscores were 0.752 and 0.696. A possible reason as to why the reliabilities are different could be because the reliability of the raw subscores was computed using coefficient alpha which is a direct measure of how much the items are correlated and the number of items in the instrument. However, reliability of the IRT subscores was computed using Equation 3.8 and it takes into account not only the variance of the subscores, but also how much uncertainty is present when estimating each of the subscores.

Finally, the Haberman Analysis for the raw scores, did not support the assumption that the raw subscores provided distinct information from the total raw score. The Haberman analysis for the raw scores can be seen in Appendix H. The correlation between each of the subscores and the total raw score was very high: 0.934 for the statistical literacy subscore and 0.944 for the statistical reasoning subscore. This supports the results from the Haberman Analysis that the subscores do

not add additional information to the total raw score. However, for the Cross-loading subscores the result was the opposite, pointing to the support for reporting the subscores.

Based on the discussion above, there is evidence to suggest that the subscores' standards set by Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) are best met for subscores from the Cross-loading IRT model. However, raw subscores using the number of items correct should not be reported since there is very small evidence that these subscores are distinct or provide distinct information from the raw total score.

The Cross-loading Model is not only useful to explain the relationship between the statistical literacy and statistical reasoning subscores, but the item characteristics estimated by this model can be used to investigate which items from the REALI assessment need to be improved. The next subsection reports information about items that presented poor item discrimination.

Item information according to the Cross-loading Model. Figure 5.6 shows the discrimination values for each of the 40 items from REALI. The items with the lowest discrimination values (smaller than 0.35) were three statistical literacy items: Item 5, Item 6, and Item 30.

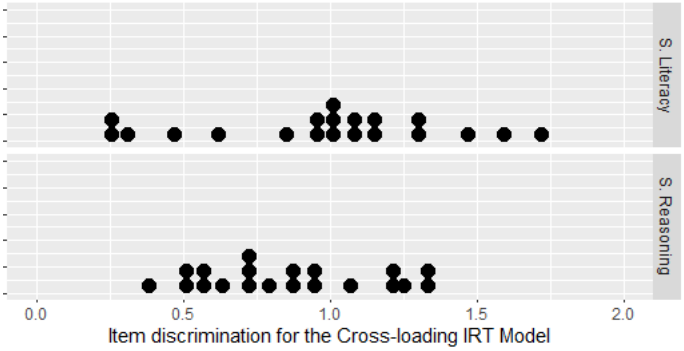


Figure 5.6. Values of discrimination for each item.

Item 30 (see Figure 5.7) was the worst discriminating item in the REALI instrument. This item was designed to assess students' ability to understand that a confidence interval for a

proportion is centered at the sample statistic. A total of 57% of the students got this item correct, but the low item discrimination gives evidence that these students were not necessarily the students with highest abilities. Around one fourth of the students with the highest abilities chose alternative B which stated that "37% of veterans in the *population* have been divorced at least once". A possible reason why this happened could be that students might be thinking of 37% as a plausible value for the population parameter since 37% is included in the confidence interval. In addition, alternative B does not state any level of confidence when making an inference about the population and students appear to not be concerned about that. This item also presented concerning results on the pilot test and even though its performance on the field test improved with the modification of the item, it seems that this item is still not performing well enough. This item will most likely be deleted and replaced by another statistical literacy item addressing the concepts of confidence interval.

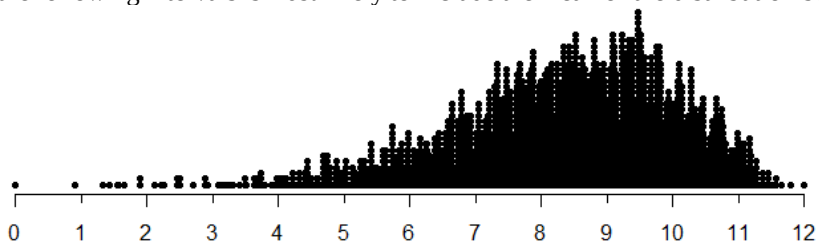
In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

a. We can say that 37% of veterans in the *sample* have been divorced at least once.
b. We can say that 37% of veterans in the *population* have been divorced at least once.
c. We can say that 95% of veterans in the *sample* have been divorced at least once.
d. We can say that 95% of veterans in the *population* have been divorced at least once.

Figure 5.7. Item 30.

Item 5 was another poorly discriminating item which was designed to assess students' ability to understand how the mean is affected by skewness. The main reason this item produced such a low discrimination was because almost all students (97%) correctly answered this item. This was the easiest item in the whole instrument and it seems to not help to differentiate students with low and high ability. This item will most likely be re-written so that the level of difficulty is increased and thus higher discrimination is achieved.

Which of the following intervals is *most* likely to include the mean of the distribution below?



a) 4 to 6
b) 7 to 9
c) 10 to 12

Figure 5.8. Item 5.

Similar to Item 5, Item 6 also belongs to the category of "Measures of Center" (see Figure 5.9). This item was designed to measure students' ability to interpret the mean in the context of the data. About half of the students correctly answered this item recognizing that the average was a summary measure representing the dogs in the sample. Alternative D behaved properly having negative discriminations of -0.20. However, alternative B and C presented discrimination values of -0.12 and -0.03. This means that some high ability students were likely to choose alternative B or alternative C as the correct answer. Alternative C gives the interpretation of the median instead of the mean, which is a common misconception. Alternative B has a similar interpretation of the average as alternative A (the correct answer), but it refers to the *population* instead of the *sample*. Students who chose alternative B might be ignoring the role of study design and making a generalization to a population even without any information about how the survey's responses was obtained. In addition, the word "national" in the stem of the item might be misleading students. Maybe, students are interpreting the word "national" to be equivalent to a *representative* sample. It seems that there is no problem with the item itself (e.g., bad item writing) and the reason for the bad discrimination could be because of students' misconceptions leading them to choose the wrong

alternatives. However, more research is needed to understand the reason why students are choosing each of the incorrect alternatives.

According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the average?

a. For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700.
b. For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700.
c. For all dog owners in this sample, about half were above $1,700 and about half were below $1,700.
d. For most owners, the first-year costs for owning a large-sized dog is $1,700.

Figure 5.9. Item 6.

5.3 Limitations

Much has been learned about the relationship between statistical literacy and statistical reasoning in this study. However, the limitations of the study are important to consider in interpreting the results. Firstly, instructors and students participated in this study on a voluntary basis, and the administration of the REALI instrument was not uniform among all institutions. Some instructors used REALI as a required part of the course, others as an extra credit opportunity or a review for the exam. Therefore, students' effort and response rate varied greatly among institutions. In addition, most if not all of the students completed the REALI assessment outside of class. This could add additional variation in students' scores due to environmental issues such as distractions. These differences in test administration might have increased error in student responses.

Another point to consider is that this instrument was composed of items covering eight areas of learning (representations of data, measures of center, measures of variability, study design, hypothesis testing and *p*-values, confidence intervals, bivariate data, and probability); however, not all students had the opportunity to learn the content presented by the items. Lack of opportunity to

learn can introduce guessing and consequently measurement error in students' responses. This adds to the uncertainty regarding students' responses, and therefore decreases the reliability of scores. Item order effects and test fatigue are also problems not measured in this study that could influence the results.

5.4 Implications for Teaching

Recommendations in the field of statistics education have been emphasizing the importance of developing statistical literacy and statistical reasoning rather than computations and procedures (e.g., GAISE-ASA, 2005; Garfield & Ben-Zvi, 2010). The results from this study supported the assumption that statistical literacy and statistical reasoning can be differentiated. Therefore, instructors need have clear statistical literacy learning goals and also statistical reasoning learning goals in their classes and show evidence, through the use of well-developed and quality assessments, that students are indeed achieving higher levels of statistical literacy and statistical reasoning. In addition, because of the hierarchy between the statistical literacy and statistical reasoning learning goals, it is important for instructors to note that statistical reasoning is not a straight forward step from statistical literacy. For instance, students need to develop a certain level of statistical literacy and then they need to be taught how to make connections and relate the different statistical concepts they learned. Therefore, instructors need to provide opportunities for students to learn how to reason with statistical concepts.

Another implication for teaching is related to assessing students at the end of a course. The REALI assessment was designed to be used with students from any type of introductory statistics course. Therefore, this instrument can be used at the end of an introductory statistics course to provide information about important statistical literacy and statistical reasoning topics to evaluate students' learning outcomes. For instance, Item 3 and Item 8 were the two hardest items in the

REALI instrument with less than 30% of the students correctly answering these items. After some investigation, it can be noticed that these two items involve graphs and the normal distribution. Further attention can be given to understand how these topics have been taught in the curriculum and why students are incorrectly answering these items. Looking at each of the distractors will give insight to possible students' misconceptions or need to curricula improvement. For instance, the curricula might be over-emphasizing the normal distribution leading the students to choose always a graph that looks more normal, without giving careful thought to the other information available in the problem. Therefore, the REALI instrument can be a tool for identifying students' misconceptions and guiding changes and improvements in statistics courses.

REALI can also be used in the evaluation of curricula or to assess the effect of curriculum changes, as long as the learning goals assessed by this instrument are closely aligned with the intended learning goals of the curricula being used in class. For instance, currently, there has been efforts to change introductory courses based on the recommendations by Cobb (2005, 2007) and *Guidelines for Assessment and Instruction in Statistics Education* (GAISE; ASA, 2005). New statistics curriculum have been developed (e.g., Garfield, delMas, Zieffler, 2012 and Tintle, VanderStoep, Holmes, Quisenberry & Swanson, 2011) using modelling and simulation approach to teaching inference. However, these new courses differ. Some courses still teach traditional content such as t-tests, on the other hand, courses such as the CATALST course (Garfield, delMas, Zieffler, 2012) do not teach traditional content; instead this course focuses on randomization tests. It is important, therefore, to examine how well students are performing on these new curricula and evaluate if there is a curriculum that is leading to better student's performance. A possible way to compare students from these different curricula is to use the REALI instrument at the end of the course and investigate if students are answering the questions in the same way or if the curricula they are in affect how students reason about statistical concepts leading to different answers.

5.5 Implications for Future Research

The results from the psychometric analysis of the data from the REALI instrument showed the existence of at least three statistical literacy items that presented low discrimination. Therefore, additional research is needed to understand why students are struggling when answering these items and to understand how these items can be improved. If items are re-written or new items are included in the instrument, it will be desired that they have a higher level of difficulty to ensure that there will be enough information to estimate students' abilities throughout all ability range.

Further research will also explore how statistical literacy and statistical reasoning are related for different types of instruction. As mentioned in the previous section, there are new introductory statistics curricula that are currently being used. These new curricula introduce the ideas of making connections between statistical concepts much earlier in the course when compared to traditional statistical courses. Therefore, research is needed to investigate if a hierarchy between statistical literacy and statistical reasoning is also present when students are learning statistics in these new type of introductory statistics courses. More research is also needed to investigate how statistical literacy and statistical reasoning relate for higher level statistics courses.

This study showed evidence that although the learning goals of statistical literacy and statistical reasoning learning goals overlap, it is possible to distinguish between them. Therefore, this research supported the theory by Garfield and Ben-Zvi (2008) and delMas (2002). Having empirical evidence to support definitions of statistical literacy and statistical reasoning and how they related will help to bring unity in statistics education research. However, the models and definitions provided by Garfield and Ben-Zvi (2008) and delMas (2002) also include the statistical thinking learning goal. Therefore, more research is needed to investigate how statistical thinking relates to statistical reasoning and statistical literacy. Statistical thinking is a very important learning goal for statistics courses and received great emphasis on the Guidelines for Assessment

and Instruction in Statistics Education (GAISE-ASA, 2005). Therefore, understanding its relationship with the other two learning goals will help guide instructors in teaching and assessing their students' ability to develop statistical thinking.

The results presented in this study also support the categorization of items by delMas (2002). He argued that it was possible to differentiate among learning goals when examining the nature of the task performed by students when answering an item. delMas (2002) stated that words such as *identify, describe, and interpret* could be used for statistical literacy items. Indeed, almost all statistical literacy items in the REALI instrument asked students to *interpret* graphs, statistical statements, confidence interval, and analysis' conclusions, *describe* hypothesis or *identify* statistics, parameters, or types of variables. The statistical reasoning questions also used the vocabulary presented by delMas (2002); however, the main characteristic of statistical reasoning items in the REALI assessment was that they involved more than one statistical concepts and forced students to make connections between these concepts. Therefore, statistical reasoning items might be more easily identified by the connections made among statistical concepts than by the type of task asked by the items. However, more research can be done through interviews with students to verify if students are indeed exhibiting statistical literacy or statistical reasoning when working through such items.

Finally, the REALI instrument administered to students in the field test was composed of only one form. Therefore, future research will examine potential item order effects and how they relate to the hierarchy between statistical literacy and statistical thinking.

5.6 Conclusion

This research study provided evidence of the distinction between the learning goals of statistical literacy and statistical reasoning. This is a valuable and significant information for the

statistics education community. This study also provided solid and research-based definitions of statistical literacy and statistical reasoning than can be used to bring unity to the research in statistics education. In addition, this study collected a strong validity argument supporting the intended inferences and uses of the subscores from the REALI instrument and its ability to measure students' statistical literacy and statistical reasoning.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Statistical Association. (2005). Guidelines for assessment and instruction in statistics education. *Published on the World Wide Web at http://www.Amstat.org/education/gaise.*

American Statistical Association. (2007), "*Using Statistics Effectively in Mathematics Education Research,"* Retrieved Nov. 02, 2013, from ASA Web site: http://www.amstat.org/education/pdfs/UsingStatisticsEffectivelyinMathEdResearch.pdf

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete edition). New York, NY: Longman.

Batanero, C., Tauber, L. M., & Sánchez, V. (2004). Students' reasoning about the normal distribution. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 257-276). Springer Netherlands.

Bakker, A. (2004b, November). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83. Retrieved July 15, 2007, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2) Bakker.pdf

Bakker, A., Biehler, R., & Konold, C. (2004). Should young students learn about Boxplots? In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education, IASE 2004 roundtable on curricular issues in statistics education*, Lund Sweden. Voorburg, The Netherlands: International Statistical Institute.

Bakker, A., & Gravemeijer, K. P.E. (2006). An historical phenomenology of mean and median. *Educational Studies in Mathematics, 62*(2), 149–168.

Ben-Zvi, D. (2004b, November). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*(2), 42–63. Retrieved July 15, 2007, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2) BenZvi.pdf

Biggs, J. (1989). Towards a Model of School-Based Curriculum Development and Assessment Using the SOLO Taxonomy. *Australian journal of education*, *33*(2), 151-63.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university*. McGraw-Hill International.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* New York: David McKay.

Breslow, N. E. (1999). Discussion: Statistical thinking in practice. *International Statistical Review,* 67(3), 252-255.

Booker, M. J. (2007). A roof without walls: Benjamin Bloom's taxonomy and the misdirection of American education. *Academic Questions*, *20*, 347-355. doi: 10.10007/s12129-007-9031-9

Boulton-Lewis, G. (1994). Tertiary students' knowledge of their own learning and a SOLO taxonomy. *Higher Education, 28*(3), 387-402.

Budgett, S., & Pfannkuch, M. (2007). Assessing students' statistical literacy. *Assessment Methods in Statistical Education: An International Perspective*, *19*, 103.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives. Handbook I: Cognitive domain. New York, NY: David McKay.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3), 1-17.

Chance, B. L., & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal*, *1*(2), 38-41.

Chance, B. L., delMas, 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). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chervaney, N., Collier, R. Fienberg, S., Johnson, P, and Neter, J. (1977), "A framework for the development of measurement instruments for evaluating the introductory statistics course," *The American Statistician*, 31, 17-23.

Chervaney, N., Benson, P.G., and Iyer, R. (1980), "The planning stage in statistical reasoning," *The American Statistician*, 34, 222-226.

Chick, H., & Pierce, R. (2013). The statistical literacy needed to interpret school assessment data. *Mathematics Teacher Education and Development, 15*(2), 5-26.

Cobb, G. (1992). Teaching statistics. *Heeding the Call for Change: Suggestions for Curricular Action, 22*, 3-43.

Cobb, G. W. (2005). The introductory statistics course: A saber tooth curriculum. In talk presented at the *United States Conference on Teaching of Statistics, Columbus, OH*.

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, 1(1). Retrieved from www.escholarship.org/uc/item/6hb 3k0nz.

Coffman, W. E. (1956). [Review of the book *Taxonomy of educational objectives. Handbook I: Cognitive domain*]. *Educational and Psychological Measurement*, 16(3), 401-405.

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education, 18*, 382–393.

delMas, R. C. (2002). Statistical literacy, reasoning and learning: A commentary. *Journal of Statistics Education, 10*(3).

delMas, R. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1-25.

Garfield, J. (1998). The statistical reasoning assessment: Development and validation of a research tool. Paper presented at the *Proceedings of the 5th International Conference on Teaching Statistics.*

Garfield, J. (1991). Evaluating students' understanding of statistics: Development of the statistical reasoning assessment. Paper presented at the *Proceedings of the Thirteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, 2,* 1-7.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, 10(3), http://www.amstat.org/publications/jse/v10n3/garfield.html

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

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for research in Mathematics Education, 19*, 44-63.

Garfield, J., & Ben-Zvi, D. (2005, May). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, *4*(1), 92–99. Retrieved December 26, 2006, from http://www.stat.auckland.ac.nz/ ~iase/serj/SERJ4(1) Garfield BenZvi.pdf

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372-396.

Garfield, J., & Ben-Zvi, D. (2008). Developing students' statistical reasoning. *Connecting Research and Teaching Practice.the Netherlands: Springer.*

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*(1-2), 99-125.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary- level statistics course. *ZDM, 44*(7), 883-898. doi:10.1007/s11858-012-0447-5

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

Garfield, J., delMas, R., & Chance, B. (n.d.). Assessment Resource Tools for Improving Statistical Thinking. Retrieved from https://app.gen.umn.edu/artist/index.html

Gould, R. (2004, November). Variability: One statistician's view. *Statistical Education Research Journal*, *3*(2), 7–16. Retrieved November 14, 2006, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2) Gould.pdf

Green, D. R. (1983). A survey of probability concepts in 3000 students aged 11–16 years. In D. R. Grey, P. Holmes, V. Barnett, & G. M. Constable (Eds.). *Proceedings of the 1st International Conference on Teaching Statistics* (pp. 766–783). Sheffield, England: Teaching Statistics Trust.

Groth, R. E. (2005). An investigation of statistical thinking in two different contexts: Detecting a signal in a noisy process and determining a typical value. *Journal of Mathematical Behavior, 24*, 109–124.

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning*, *8*, 37–63.

Hammerman, J. K., & Rubin, A. (2004, November). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17–41. Retrieved December 4, 2006, from http://www.stat.auckland.ac.nz/~iase/serj/ SERJ3(2) Hammerman Rubin.pdf

Hawkins, A., Jolliffe, F., and Glickman, L. (1992), *Teaching Statistical Concepts*, London: Longman Publishers.

Heaton, R. M., & Mickelson,W. T. (2002). The learning and teaching of statistical investigation in teaching and teacher education. *Journal of Mathematics Teacher Education, 5*, 35–59.

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing students' statistical thinking. *Mathematics Thinking and Learning, 2*, 269–307.

Jones, G. A., Langrall, C. W., Mooney, E. S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 97-117). Springer Netherlands.

Kaplan, J. J., & Thorpe, J. (2010). Post secondary and adult statistical literacy: Assessing beyond the classroom. Paper presented at the *Data and Context in Statistics Education: Towards an Evidence-Based Society. Proceedings of the Eighth International Conference on Teaching Statistics. Voorburg, the Netherlands: International Statistical Institute.*

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kahneman, D., Slovic, P. & Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Konold, C. (1989), "Informal Conceptions of Probability," *Cognition and Instruction*, *6*(1), 59-98.

Konold, C. (1990). ChancePlus: A computer-based curriculum for probability and statistics. *Final Report to the National Science Foundation Scientific Reasoning Research Institute, University of Massachusetts, Amherst.*

Konold, C. (1995), "Issues in Assessing Conceptual Understanding in Probability and Statistics," *Journal of Statistics Education, 3*(1). Retrieved from http://www.amstat.org/publications/jse/v3n1/konold.html

Konold, C., & Higgins, T. L. (2003), "Reasoning about data", In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.),*A research companion to Principles and Standards for School Mathematics*, Reston, VA: National Council of Teachers of Mathematics, pp.193-215.

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2003). *Data seen through different lenses*. Unpublished Manuscript, Amherst, MA. Retrieved September 28, 2007, from http://www.umass.edu/srri/serg/papers/Konold-Higgins, et al.pdf

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218.

Kropp, R. P., & Stoker, H. W. (1966). *The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives*. Tallahassee, FL: Institute of Human Learning and Department of Educational Research and Testing, Florida State University. (ERIC Documentation Reproduction Service No. ED 010 044)

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal, (12)*1, *20-47*. Available from: http://iase-web.org/documents/SERJ/SERJ12(1)_LaneGetaz.pdf

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics*, *23*(6), 557-568.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 149–176). Mahwah, NJ: Lawrence Erlbaum Associates.

Liu, H. J. (1998). A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States. (Unpublished doctoral dissertation, University of Minnesota).

Lovett, M. (2001), "A collaborative convergence on studying reasoning processes: A case study in statistics," *Cognition and Instruction: Twenty-Five Years of Progress*, eds. D. Klahr and S. Carver, Mahwah, NJ: Lawrence Erlbaum, 347-384.

Madaus, G. F., Woods, E. M., & Nuttall, R. L. (1973). A causal model analysis of Bloom's taxonomy. *American Educational Research Journal*, *10*(4), 253-262.

Mallows, C. (1998). The zeroth problem. *The American Statistician*, *52*(1), 1-9.

Makar, K., & Confrey, J. (2002). Comparing two distributions: Investigating secondary teachers' statistical thinking. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*. Cape Town, South Africa: International Association for Statistics Education.

Miller, W. G., Snowman, J., & O'Hara, T. (1979). Application of alternative statistical techniques to examine the hierarchal ordering in Bloom's taxonomy. *American Educational Research Journal*, *16*(3), 241-248.

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4*(1), 23–63.

Moore, D. S. (1990). Uncertainty. In L.A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-138). Washington, D. C.: National Academy Press.

Moore, D. (1999). Discussion: What shall we teach beginners?. *International Statistical Review*, 67(3), 250-252.

Moritz, J. B. (2004). Reasoning about covariation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227–256). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Nisbett, R. (1993), *Rules for Reasoning*, Mahwah, NJ: Lawrence Erlbaum.

Paul, R. W. (1993). *Critical thinking: What every person needs to survive in a rapidly changing world*. Santa Rosa, CA: Foundation for Critical Thinking.

Pfannkuch, M., & Wild, C. J. (2000). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statistical Science*, 132-152.

Pfannkuch, M., & Wild, C. J. (2003). Statistical thinking: How can we develop it? In Bulletin of the International Statistical Institute 54th Session Proceedings, Berlin, 2003 [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *Challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal*, *5*(2), 4-9.

Pring, R. (1971). Bloom's taxonomy: A philosophical critique (2). *Cambridge Journal of Education*, *1*(1), 83-91.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201–226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, *7*(2), 5-19.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3), 6-13.

Sabbag, A., & Zieffler A. (2015). Assessing Learning Outcomes: An analysis of the GOALS-2 instrument. *Statistics Education Research Journal*, *14*(2), 93–116.

Sanchez, E., & Blancarte, A.L.G. (2008). Training in-service teachers to develop statistical thinking. *Proceedings of the 8th International Conference on Teaching Statistics*.

Shaughnessy, J.M. (1992). Research in probability and statistics: Reflections and directions. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*. (pp.465-494). New York: Macmillan.

Sedlmeier, P. (1999), *Improving Statistical Reasoning: Theoretical Models and Practical Implication*, Mahwah, NJ: Lawrence Erlbaum.

Sharma, S., Doyle, P., Shandil, V., & Talakia'atu, S. (2011). Developing statistical literacy with year 9 students. *Set: Research Information for Teachers*, (1), 43.

Sinharay, S. (2010). When can subscores be expected to have added value? Results from operational and simulated data. *ETS Research Report Series*, 2010(2), i-28.

Snee, R. D. (1990). Statistical thinking and Its contribution to total quality. *The American Statistician*, 44(2), 116-121.

Snee, R. D. (1999). Discussion: Development and use of statistical thinking: A new era. *International Statistical Review*, 67(3), 255-258.

Stanley, J. C., & Bolton, D. L. (1957). [Review of the book *Taxonomy of educational objectives. Handbook I: Cognitive domain*]. *Educational and Psychological Measurement*, *17*(4), 631-634.

Tintle, N., VanderStoep, J., Holmes, V. L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1), n1.

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician, 57*(2), 74-79.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277–294). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals* (Studies in mathematical thinking and learning series). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation*, *1*(3), 247-275.

Watson, J. M., & Moritz, J. B. (2000c). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2*, 11–50.

Watson, J. M., & Moritz, J. B. (2000d). The development of concepts of average. *Focus on Learning Problems in Mathematics, 21*, 15–39.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*, 1–29.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3-46.

Wild, C. J. (2006, November). The concept of distribution. *Statistics Education Research Journal*, *5*(2), 10–26. Retrieved July 15, 2007, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2) Wild.pdf

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223-248.

Zieffler, S. A. (2006). *A longitudinal investigation of the development of college students' reasoning about bivariate data during an introductory statistics course*. Unpublished Ph.D. Thesis. University of Minnesota.

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

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education, 16*(2). Retrieved from http://www.amstat.org /publications/jse/v16n2/zieffler.pdf

Zieffler, A., & Garfield, J. (2009) Modeling the growth of students covariational reasoning during an introductory statistics course. *Statistics Education Research Journal, 8*(1), 7–31.

Ziegler, L. (2012). The effect of length of an assessment item on college student responses on an assessment of learning outcomes for introductory statistics. (Unpublished pre-dissertation paper, University of Minnesota).

Ziegler, L. (2014). Reconceptualizing statistical literacy: developing an assessment for the modern introductory statistics course (Unpublished doctoral dissertation). Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/165153.

Appendix A: Areas of learning and learning goals from each item

Item #	Instrument	Learning goals
Representation of data		
9	BLIS	Ability to describe and interpret a dotplot
10	BLIS	Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data
11	BLIS	Understanding the importance of creating graphs prior to analyzing data
17	BLIS	Understanding of what an empirical sampling distribution represents
12	GOALS	Able to reason about the type of graphic representation that is needed to represent the shape, center, and variability of the distribution of a variable.
Measures of center		
13	BLIS	Ability to interpret a mean in the context of the data
14	BLIS	Understand how a mean is affected by skewness or outliers
5	GOALS	Able to reason about factors that affect the mean and median
Measures of variability		
15	BLIS	Ability to interpret a standard deviation in the context of the data
16	BLIS	Understanding of the properties of standard deviation
24	BLIS	Understanding of how sample size affects the standard error (RAND. DIST.)
7	GOALS	Able to reason about the meaning of variability in the context of repeated measurements and in a context where small variability is desired
8	GOALS	Able to reason that given two distributions that have the same range, the one with less mass in the center has the larger standard deviation (tests for misconception that a uniform distribution has less "variability" than a non-uniform distribution)
Study design		
1	BLIS	Understanding of the difference between a sample and population
3	BLIS	Understanding that statistics computed from random samples tend to be centered at the parameter
4	BLIS	Ability to determine what type of study was conducted
5	BLIS	Ability to determine if a variable is quantitative or categorical
6	BLIS	Ability to determine if a variable is an explanatory variable or a response variable
7	BLIS	Understanding of the difference between a statistic and parameter
8	BLIS	Understanding that statistics vary from sample to sample
18	BLIS	Understanding that an empirical sampling distribution shows how sample statistics tend to vary (RANDOMIZATION DIST.)
34	BLIS	Understanding that only an experimental design with random assignment can support causal inference
35	BLIS	Understanding of the factors that allow a sample of data to be generalized to the population
1	GOALS	Able to reason about the purpose of random assignment

Item #	Instrument	Learning goals
2	GOALS	Able to reason about the factors that allow a sample of data to be representative of the population
3	GOALS	Able to reason that a correlational study design does not inference of causation
Confidence intervals		
19	BLIS	Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter
20	BLIS	Understanding that a confidence interval provides plausible values of the population parameter
21	BLIS	Understanding that a confidence interval for a proportion is centered at the sample statistic
22	BLIS	Understanding of how the confidence level affects the width of a confidence interval
13	GOALS	Able to reason about a misinterpretation of a confidence level (using it to make a prediction for a single case)
19	GOALS	Able to identify the 95% confidence interval for a context given a point estimate of the population parameter and a graphic representation from which the sampling variability can be estimated
11	GOALS	Able to reason about how the width of a confidence interval is related to sample size.
Hypothesis testing & *p*-values		
23	BLIS	Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis
25	BLIS	Understanding that a randomization distribution tends to be centered at the hypothesized null value
26	BLIS	Ability to estimate a p-value using a randomization distribution
27	BLIS	Understanding of the logic of a hypothesis test
28	BLIS	Understanding of the purpose of a hypothesis test
29	BLIS	Ability to determine a null and alterative hypothesis statement based on a research question
30	BLIS	Ability to determine a null and alterative hypothesis statement based on a research question
31	BLIS	Ability to determine statistical significance based on a p-value
32	BLIS	Understanding that errors can occur in hypothesis testing
33	BLIS	Understanding of how a significance level is used to make decisions
6	GOALS	Able to reason that a large p-value does not provide significant evidence of an effect.
9	GOALS	Able to reason about how differences in variability affect strength of evidence against the null hypothesis of no difference
14	GOALS	Able to reason that a smaller p-value provides stronger evidence against the null hypothesis than a larger p-value
15	GOALS	Able to reason about what the null model represents in a research study

Item #	Instrument	Learning goals
16	GOALS	Able to reason about what model should be used for the null hypothesis when comparing two groups
17	GOALS	Able to reason about a conclusion based on a statistically significant *p*-value in the context of a research study that compares two groups
18	GOALS	Able to reason about an incorrect interpretation of a p-value (probability of a treatment being more effective)
20	GOALS	Able to reason about how increasing the sample size affects the p-value, all else being equal

Appendix B: Behaviors for answering the items correctly

B1 - Behaviors for the items in the BLIS assessment.

Topic: Data Production
Learning Outcome: Understanding of the difference between a sample and population

1. The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population.

    a. The population is all American adults in 2013. The sample is the 21% of American adults that have had an email or social networking account compromised.
    b. The population is the 1,002 American adults surveyed. The sample is all American adults in 2013.
    c. The population is all American adults in 2013. The sample is the 1,002 American adults surveyed.

Behavior

- To answer this question correctly, students need to
    - Know the difference between *sample* and *population*.
    - identify the sample and the population from a reading passage (interpretation)

Topic: Data Production
Learning Outcome: Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term

2. Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 50 times and the other student flips a coin 100 times. Which student is more likely to get 48% to 52% of their coin flips heads up?

    a. The student who flips the coin 50 times because the percent that are heads up is less likely to be exactly 50%.
    b. The student who flips the coin 100 times because that student has more chances to get a coin flip that is heads up.
    c. The student who flips the coin 100 times because the more flips that are made will increase the chance of approaching a result of 50% heads up.
    d. Neither student is more likely because the flipping of the coin is random and therefore you cannot predict the outcome of the flips.

Behavior

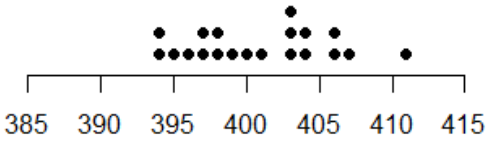
- To answer this question correctly, students need to
    - Know that in the long run the probability of the number of heads gets closer to 50% than in the short run.

- Know that the percentage of heads does not change from flip to flip.
- Know that even if the process is random, you can still see a pattern in the long run and make predictions.

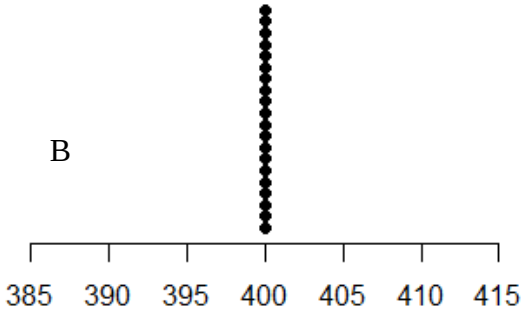
Topic: Data Production
Learning Outcome: Understanding that statistics computed from random samples tend to be centered at the parameter

3. A manufacturer of frozen pizzas produces sausage pizzas, which are intended to have an average weight of 400 grams. To check the quality of the manufacturing process, random samples of 25 pizzas are taken daily and the average weight of the pizza's in the sample is recorded. Assuming that nothing is wrong with the manufacturing process, which of the following graphs is the most plausible for the average weight in each of the 20 samples?

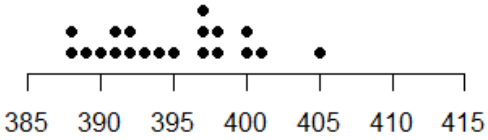
A



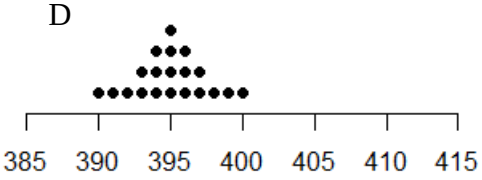
B



C



D



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

Behavior
- To answer this question correctly, students need to
  - Know that 400 is the population parameter.
  - Know what are random samples
  - Know that that statistics from random samples tend to be centered at the population parameter
  - Students need to recognize that answer c and d are not centered at 400.
  - Know that statistics from different samples will vary

Topic: Data Production
Learning Outcome: Ability to determine what type of study was conducted

4. Suppose a researcher wanted to determine if aspirin reduces the chance of a heart attack. The researcher studied 500 patients who visited a regional hospital in the last year. Half (250) of the patients were randomly assigned to take aspirin every day and the other half to take a placebo everyday. Then after a certain length of time, the percentage of heart attacks for the patients who took aspirin every day and for those who did not take aspirin every day were reported. What type of study did the researcher conduct?

    a. Observational
    b. Experimental
    c. Survey

Behavior
- To answer this question correctly, students need to
    - Know what an observational study is.
    - Know what an experimental study is.
    - Know what a survey is.
    - Identify the type of study based on a description of the study.

Items 5 and 6 refer to the following situation:
A student is gathering data on the driving experiences of other college students. One of the variables measured is the type of car the student drives. These data are coded using the following method: 1 = compact, 2 = subcompact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.

Topic: Data Production
Learning Outcome: Ability to determine if a variable is quantitative or categorical

5. What type of variable is this?

    a. Categorical
    b. Quantitative
    c. Continuous

Behavior
- To answer this question correctly, students need to
    - Know what a categorical variable is.
    - Know what a quantitative variable is.
    - Know what a continuous variable is.
    - Identify the type of variable based on a description.

Topic: Data Production
Learning Outcome: Ability to determine if a variable is an explanatory variable or a response

variable

6. The student plans to see if the type of vehicle a student drives is a predictor of the number of speeding tickets he or she gets in a year. Identify the response variable in this study.

    a. College students
    b. Type of vehicle
    c. Number of speeding tickets
    d. Average number of speeding tickets last year

Behavior

- To answer this question correctly, students need to
    - Know what a response variable is.
    - Know what a predictor is.
    - Identify what is the response variable based on a description.

Topic: Data Production
Learning Outcome: Understanding of the difference between a statistic and parameter

7. CNN conducted a quick vote poll with a random sample of 5,581 Americans on September 19, 1999 to determine "What proportion of Americans still think that the Miss America pageant is still relevant today?" For the sample, 1,192 people voted yes and 4,389 people voted no. Identify the statistic and parameter of interest.

    a. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the 5,581 Americans who took part in the survey.
    b. The statistic is the 5,581 Americans who took part in the survey and the parameter is all Americans.
    c. The statistic is the proportion of all Americans who think the pageant is still relevant and the parameter is the sample proportion of people who voted yes (1192/5581= .214).
    d. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the proportion of all Americans who think the pageant is still relevant.

Behavior

- To answer this question correctly, students need to
    - What a parameter is and that it is related to the population of interest.
    - What a statistic is and that it is related to the sample.
    - Identify the statistic and parameter from a description of the study.

Topic: Data Production
Learning Outcome: Understanding that statistics vary from sample to sample

8. In a study, Researcher A took a random sample of 25 college students and found the mean number of times they went out to eat during the last week was 4.1. In another study, Researcher B took a random sample of 25 students from the same college and

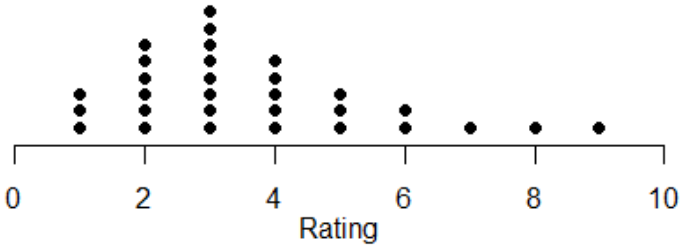
found the mean number of times they went out to eat during the last week was 3.7. What is the best explanation for why the samples taken by Researcher A and Researcher B did not produce the same mean?

a. The sample means varied because they are small samples.
b. The sample means varied because the samples were not representative of all college students.
c. The sample means varied because each sample is a different subset of the population.

Behavior
- To answer this question correctly, students need to
    - Know that different samples will give different estimates.
    - Know that the previous statement is still true if the sample size is large.
    - Know that no matter if a sample is representative or not, different samples will give different estimates.

Topic: Graphs
Learning Outcome: Ability to describe and interpret a dotplot

9. One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. Below is the distribution of this variable for the 30 students in the class.



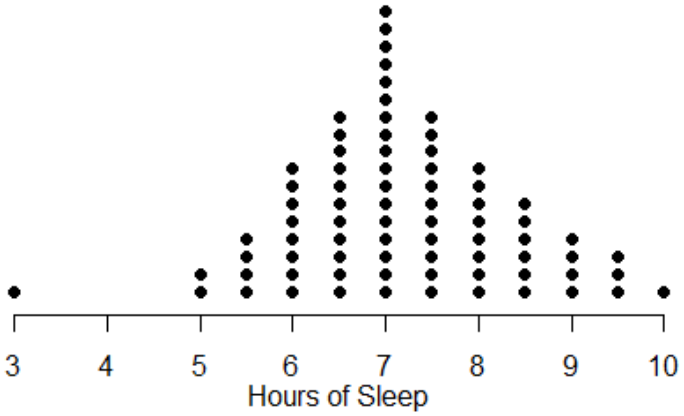
How should the instructor interpret the students' perceptions regarding their success in the class?

a. A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
b. A majority of students in the class rated their confidence as a 3 although some ratings were higher and some ratings were lower.
c. A majority of students will not try to do well in the course because they do not feel that they will succeed in statistics.

Behavior
- To answer this question correctly, students need to
    - Know how to interpret a data plot.

Topic: Graphs
Learning Outcome: Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data

10. The following graph shows the distribution of hours slept the previous night by a group of college students.



Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.

a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five.
b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
c. Many students seem to be getting 7 hours of sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.
d. The distribution of hours of sleep is somewhat normal, with an outlier at 3. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.

Behavior
- To answer this question correctly, students need to
    - Know how to interpret a data plot.
    - Know how to estimate the mean from a dotplot.
    - Know how to estimate the standard deviation from a dotplot.

Topic: Graphs
Learning Outcome: Understanding the importance of creating graphs prior to analyzing data

11. A researcher was interested in determining if a smaller dose of a particular flu vaccine was just as successful as a full dose. An experiment was conducted with 100 participants. Half of the participants were randomly assigned to receive the full dose of the vaccine and the other half received a half dose of the vaccine. The number of days the participant had flu symptoms during the following year was recorded. The researcher plans to conduct a hypothesis test to determine if there is a significant

difference in the average number of days participants had flu symptoms for the full dose group and half dose group. Which of the following is a reason why the researcher should create and examine graphs of the number of days participants had flu symptoms before the hypothesis test is conducted?

a. To decide what the null hypothesis and alternative hypothesis should be.
b. To compute the average number of days participants had flu symptoms in order to conduct a hypothesis test.
c. To see if there are recognizable differences in the two groups to decide if a hypothesis test is necessary.

Behavior
- To answer this question correctly, students need to
    - Know why is it important to create graphs prior to analyzing the data.
    - Know that the null and alternative hypothesis comes from research ideas and not from the descriptive analysis.

Topic: Descriptive statistics
Learning Outcome: Ability to interpret a probability in the context of the data

12. According to the National Cancer Institute, the probability of a man in the United States developing prostate cancer at some point during his lifetime is .15. What does the statistic, .15, mean in the context of this report from the National Cancer Institute?

a. For all men living in the United States, approximately 15% will develop prostate cancer at some point in their lives.
b. If you randomly selected a male in the United States there is a 15% chance that he will develop prostate cancer at some point in his life.
c. In a random sample of 100 men in the United States, 15 men will develop prostate cancer.
d. Both a and b are correct.

Behavior
- To answer this question correctly, students need to
    - Know how to interpret a probability given a context.

Topic: Descriptive statistics
Learning Outcome: Ability to interpret a mean in the context of the data

13. According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the mean?

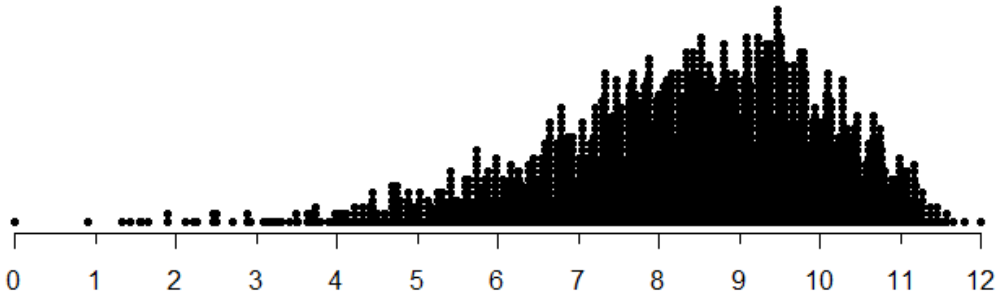
a. For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700.
b. For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700.

c. For all dog owners in this sample, about half were above $1,700 and about half were below $1,700.
d. For most owners, the first-year costs for owning a large-sized dog is $1,700.

Behavior
- To answer this question correctly, students need to
  - Know how to interpret a mean given a context.

Topic: Descriptive statistics
Learning Outcome: Understand how a mean is affected by skewness or outliers

14. The distribution for a population of measurements is presented below.



A sample of 10 randomly selected values will be taken from the population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?

a. 6 to 7
b. 8 to 9
c. 9 to 10
d. 10 to 11

Behavior
- To answer this question correctly, students need to
  - Know how to estimate the mean from a dotplot, in which the distribution is left skewed.

Topic: Descriptive statistics
Learning Outcome: Ability to interpret a standard deviation in the context of the data

15. Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quiz scores was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?

a. All of the individual scores are one point apart.
b. The difference between the highest and lowest score is 1 point.
c. The difference between the upper and lower quartile is 1 point.
d. A typical distance of a score from the mean is 1 point.

Behavior
- To answer this question correctly, students need to
    - Know how to interpret a standard deviation given a context

Topic: Descriptive statistics
Learning Outcome: Understanding of the properties of standard deviation

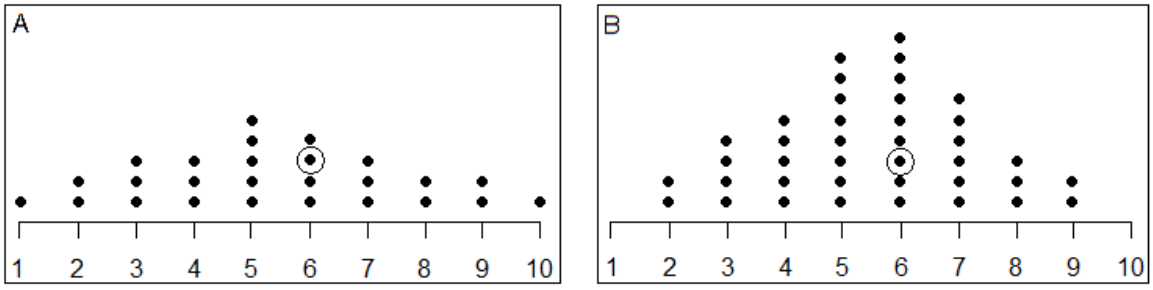
16. A teacher gives a 15-item science test. For each item, a student receives one point for a correct answer; 0 points for no answer; and loses one point for an incorrect answer. Total test scores could range from -15 points to +15 points. The teacher computes the standard deviation of the test scores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

Behavior
- To answer this question correctly, students need to
    - Know that the standard deviation is always a positive number.

Topic: Empirical sampling distributions
Learning Outcome: Understanding of what an empirical sampling distribution represents

17. Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights for 39 random samples of 3 pebbles each, with all mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.



    a. No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
    b. Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6.
    c. Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B represents the average weight of 3 pebbles.

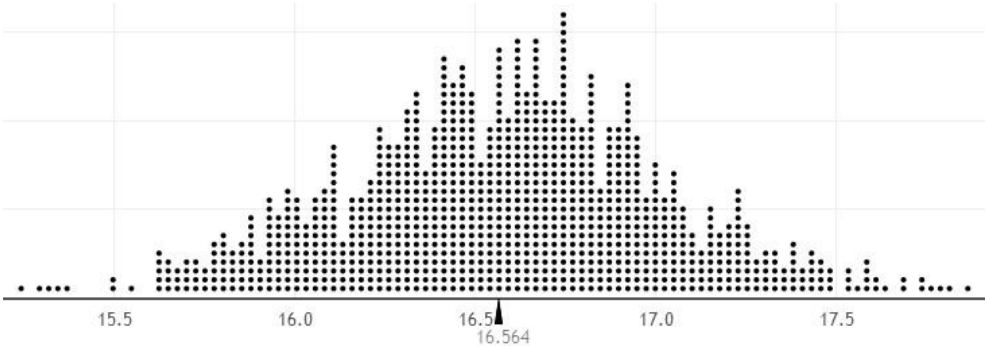
Behavior

- To answer this question correctly, students need to
    - Know how to interpret a dotplot given a context.

Items 18 and 19 refer to the following situation:
The Pew Research Center surveyed 2,986 adults chosen at random in 2011 and asked "During the past 12 months, about how many books did you read either all or part of the way through?" The sample average number of books read was 16.6. An empirical sampling distribution was estimated by doing the following:

- From the original sample, 2,986 adults were chosen randomly, with replacement.
- The mean was computed for the new sample and placed on the plot shown below.
- This was repeated 999 more times.

Below is the plot of the estimated empirical sampling distribution. (Note: This plot can be used as an estimate of what the sampling distribution would look like.)



Mean Number of Books

Topic: Empirical sampling distributions
Learning Outcome: Understanding that an empirical sampling distribution shows how sample statistics tend to vary

18. Which of the following is the best description of the variability in the empirical sampling distribution?

    a. The mean number of books that adults read during the year of 2011 was 16.564.
    b. The variability in the mean number of books from sample to sample is quite small spanning from approximately 15 to 18.
    c. The variability in the number of books from person to person is quite small spanning from approximately 15 to 18.

Behavior
- To answer this question correctly, students need to
    - Know what is an EMPIRICAL sampling distribution
    - Understand the simulation (bootstrap) process
    - Understand that each dot represents a possible MEAN not and individual value.

- Differentiate description of center form description of variability

Topic: Empirical sampling distributions
Learning Outcome: Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter

19. What values do you believe would be LESS plausible estimates of the population average number of books read if you wanted to estimate the population average with 95% confidence?

    a. Values approximately 17.2 and above because it is unlikely that adults would read that many books.
    b. Values below approximately 15.0 and values above approximately 18.0 because there are no dots that are that extreme.
    c. Values in the bottom 5% (below approximately 16.0) and values in the top 5% (above approximately 17.0).
    d. Values in the bottom 2.5% (below approximately 15.7) and values in the top 2.5% (above approximately 17.4).

Behavior
- To answer this question correctly, students need to
    - Know how to compute a 95% confidence interval from a sampling distribution. (Leave 2.5% in each tail)
    - Know that statistics in the tails of the sampling distributions are not plausible estimate of a population parameter

Topic: Confidence Intervals
Learning Outcome: Understanding that a confidence interval provides plausible values of the population parameter

20. The Pew Research Center surveyed 2,076 American adult cell phone users chosen at random in 2013. The sample percent of adult cell phone users who access the internet on their phones was 60%. The 95% confidence interval was 58% to 62%. What is this interval attempting to estimate?

    a. The average number of American adult cell phone users who access the internet on their phones in 2013.
    b. The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
    c. The percent of all American adult cell phone users who access the internet on their phones in 2013.
    d. For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.

Behavior
- To answer this question correctly, students need to
    - Recognize that the parameter which is trying to be estimated is a percentage

and not a mean.
    - Recognize what is the population related to this interval.
    - Understand that the level of confidence used to compute the CI is related to how confident we are that the interval contains the population parameter.

Topic: Confidence Intervals
Learning Outcome: Understanding that a confidence interval for a proportion is centered at the sample statistic

21. In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

    a. We know that 37% of veterans in the *sample* have been divorced at least once.
    b. We know that 37% of veterans in the *population* have been divorced at least once.
    c. We can say with 95% confidence that 37% of veterans in the *sample* have been divorced at least once.
    d. We can say with 95% confidence that 37% of veterans in the *population* have been divorced at least once.

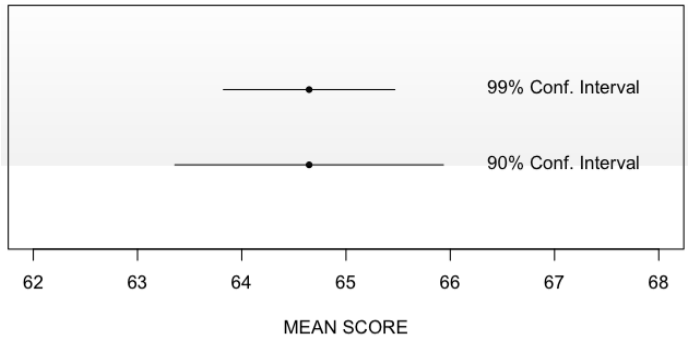
Behavior
- To answer this question correctly, students need to
    - Know that 37% is the sample statistic.
    - Know that the sample statistic is located at the center of the confidence interval.

Topic: Confidence Intervals
Learning Outcome: Understanding of how the confidence level affects the width of a confidence interval

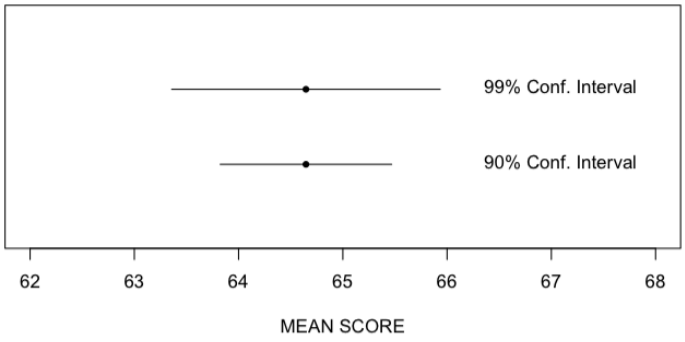
22. Consider a standardized test that has been given to thousands of high school students.

Imagine that a random sample of n = 100 test scores is drawn from the thousands of test scores. A 99% confidence interval for the population mean *and* a 90% confidence interval for the population mean are constructed using this new sample.
For the following options, a confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval. Which of the options would best represent how the two confidence intervals would compare to each other?

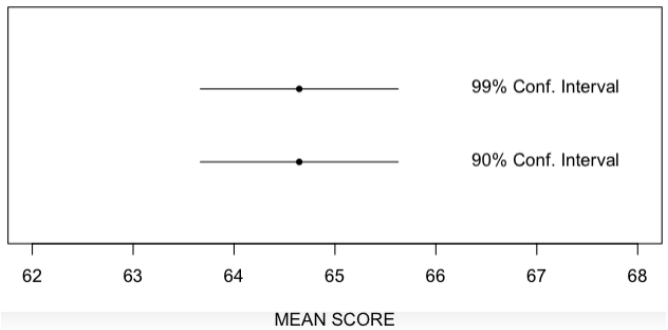
a.



b.



c.



Behavior

- To answer this question correctly, students need to
  - know how the confidence interval is affected by the level of confidence:

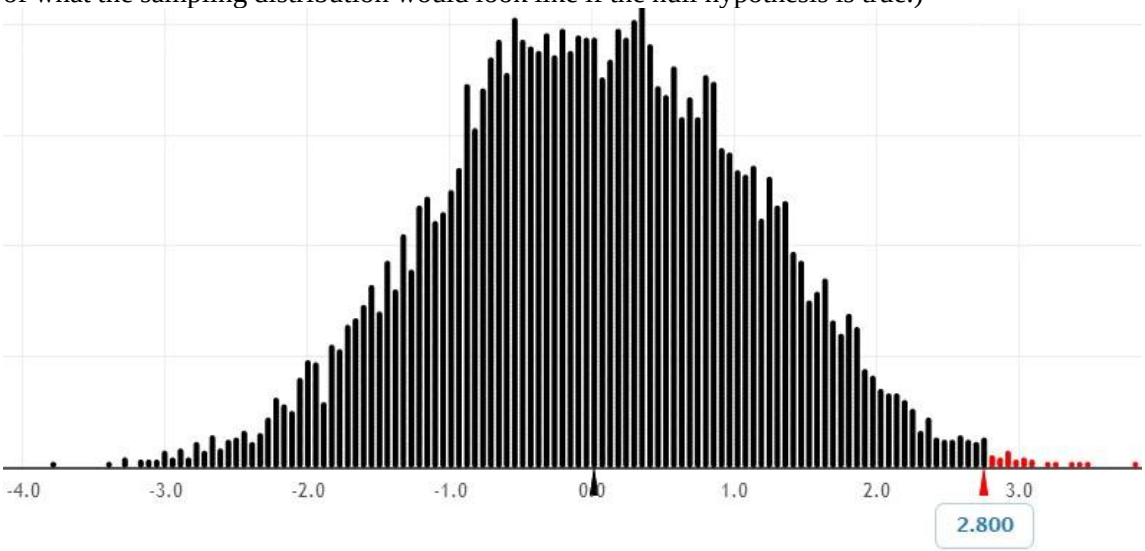
more confidence → broader interval.

Items 23 and 24 refer to the following situation:
Are people able to recall words better after taking a nap or taking a caffeine pill? A randomized experiment was conducted with 24 participants. Participants were shown a list of words in the morning. In the afternoon, half of the participants were randomly assigned to take a nap and the other half took a caffeine pill. The response variable was the number of words participants were able to recall 7 hours after being shown the list of words in the morning. The nap group recalled an average of 15.8 words and the caffeine group recalled an average of 13.0 words, with a mean difference of 15.8 – 13.0 = 2.8 words.

A randomization distribution was produced by doing the following:

- From the original sample, the 24 participants were re-randomized to the nap group (n=12) or caffeine group (n=12), without replacement.
- The mean difference in words recalled between the two re-randomized groups was computed [mean(nap group – mean(caffeine group)] and placed on the plot shown below.
- This was repeated 999 more times.

Below is the plot of the randomization distribution. (Note: This plot can be used as an estimate of what the sampling distribution would look like if the null hypothesis is true.)



(Mean Words Recalled for Nap Group) – (Mean Words Recalled for Caffeine Group)

Topic: Randomization distributions
Learning Outcome: Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis

23. The null hypothesis is there is no difference in the true mean number of words recalled

for the nap group and caffeine group. Looking at the observed sample mean difference in number of words recalled between the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis?

a. No, because the average of the re-randomized sample mean differences is equal to 0.
b. No, because the proportion of re-randomized sample mean differences equal to or above 2.8 is very small.
c. Yes, because the proportion of re-randomized sample mean differences equal to or above 2.8 is very small.
d. Yes, because the observed result shows that the nap group remembered an average of 2.8 words more than the caffeine group.

Behavior
- To answer this question correctly, students need to
    - Know how to compute evidence to reject or fail to reject the null hypothesis (*p*-value)
    - Recognize that a small p-value gives you evidence against the null model.
    - Know that the center of the sampling distribution in on zero because this distribution is based on the no difference (null) hypothesis. Exclude this option!
    - Know that the observed result is not enough to say if there is evidence or not against the null.

Topic: Randomization distributions
Learning Outcome: Understanding of how sample size affects the standard error

24. Suppose the sample size was doubled from 24 participants to 48 participants and the participants were still randomly assigned into two groups of equal size. How would you expect the standard error of the mean difference to change?

a. Decrease, because with a larger sample size, there would be less variability in the re-randomized sample mean differences.
b. Increase, because with a larger sample size, there is more opportunity for error.
c. Stay about the same, because people are still being assigned to groups randomly.

Behavior
- To answer this question correctly, students need to
    - know how the standard error is affected by the sample size. Specifically that a larger sample size leads to a decrease in standard error.

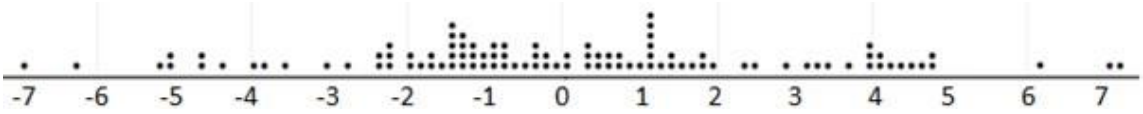
Items 25 and 26 refer to the following situation:
An experiment was conducted with 50 obese women. All women participated in a weight loss program. Twenty-six women were randomly assigned to receive daily text messages which asked participants to report their performance from the previous day. The remaining 24 women were in the control group that did not receive text messages. The average weight loss was 2.8 pounds for the text message group and -2.6 pounds for the control group. Note that the control

group had a negative average weight loss which means that they actually gained weight, on average. The difference in average weight loss was (2.8) – (-2.6) = 5.4 pounds.

A randomization distribution was produced by doing the following:

- From the original sample, the 50 women were re-randomized to the text message group (n=26) or control group (n=24), without replacement.
- The mean difference in weight loss between the two re-randomized groups was computed [mean(text message) – mean(control)] and placed on the plot shown below.
- This was repeated 99 more times.

Below is the plot of the randomization distribution for the 100 simulated mean differences. (Note: This plot can be used as an estimate of what the sampling distribution would look like if the null hypothesis is true.)



(Mean Weight Loss for Text Message Group) – (Mean Weight Loss for Control Group)
Topic: Randomization distributions
Learning Outcome: Understanding that a randomization distribution tends to be centered at the hypothesized null value

25. Why is the randomization distribution centered at 0?

a. Because the randomization distribution was created under the assumption of a difference in mean weight loss of 0.
b. Because the women who gained weight cancelled out the women who lost weight resulting in a mean of 0.
c. Because that was the original weight loss that participants started at for both groups.

Behavior
- To answer this question correctly, students need to
    - know that a randomization distribution tends to be centered at the hypothesized null value

Topic: Hypothesis tests
Learning Outcome: Ability to estimate a p-value using a randomization distribution

26. Researchers hypothesize that text messages lead to more weight loss than no text messages for women participating in this weight loss program. Compute the approximate *p*-value for the observed difference in mean weight loss of 5.4 based on the randomization distribution using the one-tailed test appropriate to the researchers' hypothesis.

    a. .03
    b. .05
    c. .06

Behavior
- To answer this question correctly, students need to
    - Know how to calculate a p-value for a 1 tailed distribution based on the sampling distribution.

Topic: Hypothesis tests
Learning Outcome: Understanding of the logic of a hypothesis test

27. The following situation models the logic of a hypothesis test. An electrician tests whether or not an electrical circuit is good. The null hypothesis is that the circuit is good. The alternative hypothesis is that the circuit is not good. The electrician performs the test and decides to reject the null hypothesis. Which of the following statements is true?

    a. The circuit is definitely not good and needs to be repaired.
    b. The circuit is most likely not good, 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 might not be good.

Behavior
- To answer this question correctly, students need to
    - Know that even if the null hypothesis is rejected, that does not mean that we are absolutely certain that the null is false.

Topic: Hypothesis tests
Learning Outcome: Understanding of the purpose of a hypothesis test

28. A researcher wants to answer the following research question: Do sleep apnea patients look more attractive after receiving a new treatment (continuous positive airway pressure therapy)? Twenty patients with sleep apnea were given the new treatment. Pictures of the patients were taken before the treatment as well as two months after the treatment. An independent rater was shown the before and after pictures for each patient in a random order and asked to determine which picture the patient looked more attractive in. Seventy percent (70%) of the patients were rated as more attractive in the

after picture. Is there a need to conduct a hypothesis test to determine whether the treatment is effective, or could the researcher just use the sample statistic (70%) as evidence of the effectiveness?

a. The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%.
b. The researcher should conduct a hypothesis test because a hypothesis test is always appropriate.
c. The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.

Behavior
- To answer this question correctly, students need to
    - Know that the observed result is not enough to say if there is evidence or not that the treatment is working since the results could have happened by chance.

Items 29 and 30 refer to the following situation:
The Amherst H. Wilder Foundation conducts surveys with homeless people in Minnesota. In their 2012 survey of 4,465 homeless adults, one question asked was "In the last 30 nights, how many nights have you spent outside, in a vehicle or vacant building, or some other place not intended for housing?" One research question that the surveyors had was "Is there a difference between males and females with regards to the average number of nights spent in a place not intended for housing?"

Topic: Hypothesis tests
Learning Outcome: Ability to determine a null and alterative hypothesis statement based on a research question

29. Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question?

a. There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
b. There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
c. There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
d. There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.

Behavior
- To answer this question correctly, students need to
    - Know what the null hypothesis is based on a context.

Topic: Hypothesis tests

Learning Outcome: Ability to determine a null and alterative hypothesis statement based on a research question

30. Which of the following is a statement of the alternative hypothesis for a statistical test designed to answer the research question?

    a. There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
    b. There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
    c. There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
    d. There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.

Behavior

- To answer this question correctly, students need to
    - Know what the alternative hypothesis is based on a context.

Topic: Hypothesis tests
Learning Outcome: Ability to determine statistical significance based on a p-value

31. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain?

    a. A large *p*-value.
    b. A small *p*-value.
    c. The magnitude of a *p*-value has no impact on statistical significance.

Behavior

- To answer this question correctly, students need to
    - Know that a small p-value gives you evidence against the null and that means that the results are statistically significant.

Topic: Hypothesis tests
Learning Outcome: Understanding that errors can occur in hypothesis testing

32. A clinical trial was conducted to determine if women who have regular mammograms to screen for breast cancer would decrease breast cancer mortality. The null hypothesis is women who have regular mammograms have the same breast cancer mortality rate as women who do not have regular mammograms. The alternative hypothesis is women who have regular mammograms have a lower breast cancer mortality rate than women who do not have regular mammograms. A hypothesis test was conducted and the results were not statistically significant. Does that mean that the null hypothesis is true, that women who have regular mammograms have the same breast cancer mortality rate as women who do not have regular mammograms?

a. Yes. It means you cannot conclude that the alternative hypothesis is true, so the null hypothesis must be true.
b. No. It means you cannot conclude that the null hypothesis is true, so the alternative hypothesis must be true.
c. No. It means that there is not enough evidence to conclude that the null hypothesis is false.
d. No. It means that there is not enough evidence to conclude that the alternative hypothesis is false.

Behavior
- To answer this question correctly, students need to
    - Know that results that are not significant means that the null hypothesis was not rejected.
    - Know that failing to reject the null hypothesis does not mean that the null is true

Topic: Hypothesis tests
Learning Outcome: Understanding of how a significance level is used to make decisions

33. Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up different scents. An experiment was conducted with a dog in Japan who was trained to smell bowel cancer in stool samples. In a test, the dog was presented with five stool samples; one from a cancer patient and four from healthy people. The dog indicated which stool sample was from the cancer patient. This was repeated a total of 38 times. Out of the 38 tests, the dog correctly identified the cancer sample 37 times. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies cancer more than one fifth of the time. The p-value is less than .001. Assuming it was a well- designed study, use a significance level of .05 to make a decision.

    a. Reject the null hypothesis and conclude that the dog correctly identifies cancer more than one fifth of the time.
    b. There is enough statistical evidence to prove that the dog correctly identifies cancer more than one fifth of the time.
    c. Do not reject the null hypothesis and conclude there is no evidence that the dog correctly identifies cancer more than one fifth of the time.

Behavior
- To answer this question correctly, students need to
    - Use the p-value to make a conclusion about the problem.

Topic: Scope of conclusions
Learning Outcome: Understanding that only an experimental design with random assignment can support causal inference

34. A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level?

a. Observational study
b. Randomized experiment
c. Survey

Behavior
- To answer this question correctly, students need to
    - Know what an observational study is.
    - Know what a randomized experiment is.
    - Know what a survey is.
    - Know which study design allows for a causation conclusion.

Topic: Scope of conclusions
Learning Outcome: Understanding of the factors that allow a sample of data to be generalized to the population

35. A college official conducted a survey to estimate the proportion of students currently living in dormitories about their preference for single rooms, double rooms, or multiple (more than two people) rooms in the dormitories on campus. A random sample of 500 first-year students was selected and the official received survey results from 160 of these students.

Which of the following does **NOT** affect the college official's ability to generalize the survey results to all dormitory students at this college?

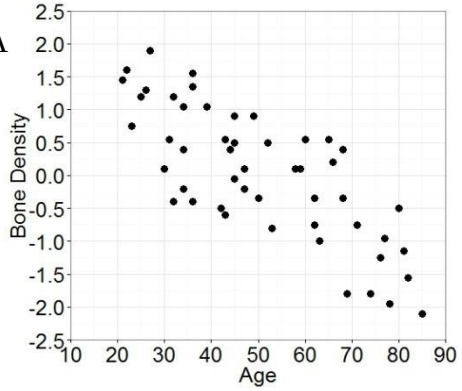
a. Although 5,000 students live in dormitories on campus, only 500 were sent the survey.
b. The survey was sent to only first-year students.
c. Of the 500 students who were sent the survey, only 160 responded.
d. All of the above present a problem for generalizing the results to all dormitory students at this college.

Behavior
- To answer this question correctly, students need to
    - Know that the results from a random sample are generalizable to the population even if the sample size is not very big.
    - Know what type of bias can happen in sampling

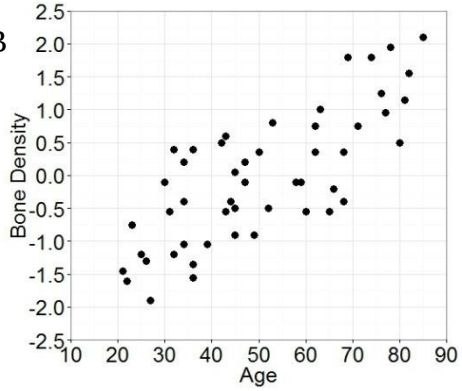
Topic: Regression and correlation
Learning Outcome: Ability to match a scatterplot to a verbal description of a bivariate relationship

36. Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?

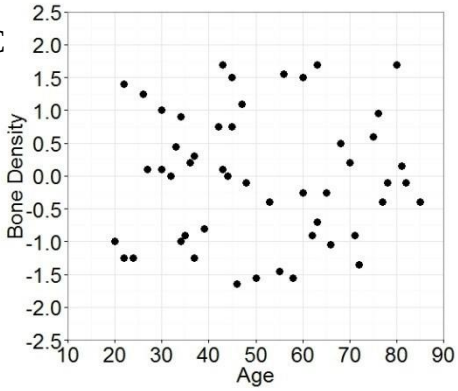
A



B



C

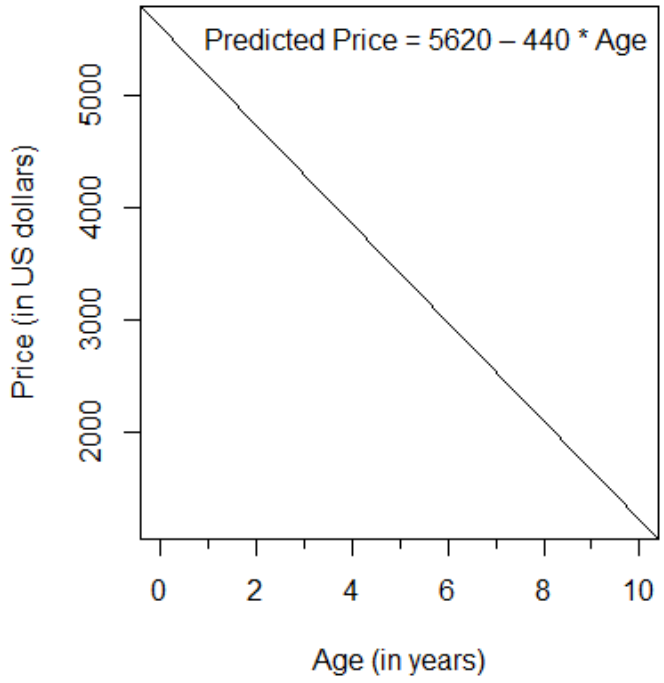


a. Graph A
b. Graph B
c. Graph C

Behavior
- To answer this question correctly, students need to
    - Know how to interpret a scatterplot.
    - Know how to match a scatterplot to a verbal description of a bivariate relationship

Topic: Regression and correlation
Learning Outcome: Ability to use a least-squares regression equation to make a prediction

37. A statistics student gathered data on a large numbers of cars of a particular model, from new cars to those that were up to 10 years old. Using the data on car ages (in years) and car prices (in US dollars) he found a linear relationship and produced the following regression equation and plot of the regression equation:



A friend asked him to use regression to predict the price of a 5 year-old model of this car. Which of the following methods can be used to provide an estimate?

a. Locate the point on the line that corresponds to an age of 5 and read off the corresponding value on the y axis.
b. Substitute an age of 5 in the equation and solve for "Predicted Price".
c. Both of these methods are correct.
d. Neither of these methods is correct.

Behavior

- To answer this question correctly, students need to
    - Know how to make a prediction from a regression line

B2 - Behaviors for the items in the GOALS assessment.

Area of reasoning: (1) Reasoning about the role of study design in drawing a statistical inference.
Learning goal: Able to reason about the purpose of random assignment

1. A research study randomly assigned participants into two groups. One group was given Vitamin E to take daily. The other group received only a placebo pill. The research study followed the participants for eight years. After the eight years, the proportion of each group that developed a particular type of cancer was compared.

What is the primary reason that the study used random assignment?

a. To ensure that the groups are similar in all respects except for the level of Vitamin E.
b. To ensure that a person doesn't know whether or not they are getting the placebo.
c. To ensure that the study participants are representative of the larger population.

Behavior

- To answer this question correctly, students need to
    - Know what random assignment is.
    - Know what the purpose of random assignment is.
    - Distinguish random assignment from random sampling and the inferences that can be made based on the study design.

Area of reasoning: (1) Reasoning about the role of study design in drawing a statistical inference.
Learning goal: Able to reason about the factors that allow a sample of data to be representative of the population.

2. A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.
Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

a. Valid, because the sample size is large enough to represent the population.
b. Valid, because 67% is far enough above 50% to predict a majority vote.
c. Invalid, because the sample is too small given the size of the population.
d. Invalid, because the sample may not be representative of the population.

Behavior

- To answer this question correctly, students need to
    - Know that a random sample is needed to allow for generalization of the results

- Understand that the size of a sample is not related to the representativeness of the sample.

Area of reasoning: (1) Reasoning about the role of study design in drawing a statistical inference.
Learning goal: Able to reason that a correlational study design does not inference of causation.

3. Researchers conducted a survey of 1,000 randomly selected adults in the United States and found a strong, positive, statistically significant correlation between income and the number of containers the adults reported recycling in a typical week.
Can the researchers conclude that higher income causes more recycling among U.S. adults? Select the best answer from the following options.

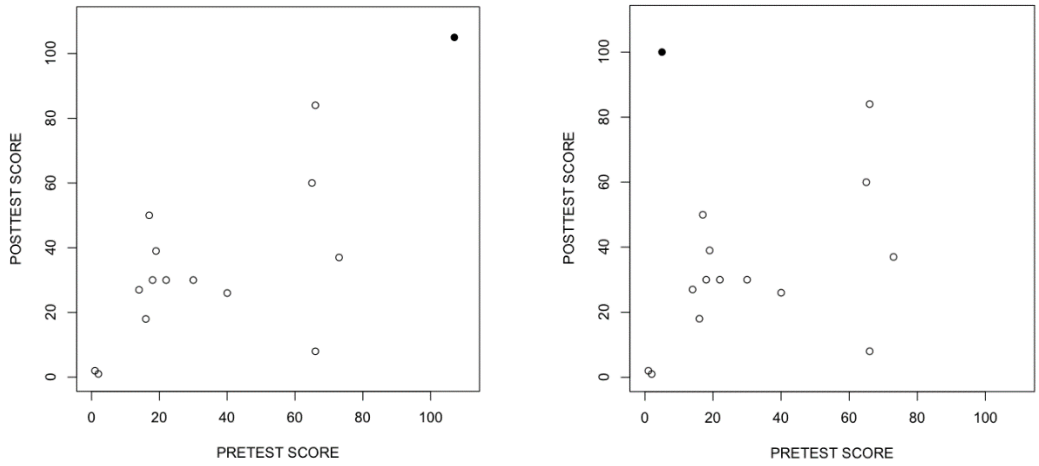
a. No, the sample size is too small to allow causation to be inferred.
b. No, the lack of random assignment does not allow causation to be inferred.
c. Yes, the statistically significant result allows causation to be inferred.
d. Yes, the sample was randomly selected, so causation can be inferred.

Behavior
- To answer this question correctly, students need to
  - Know that without random assignment in the design of the study, it is not proper to make causal conclusions.
  - Distinguish random assignment from random sampling and the inferences that can be made based on the study design. (same thing as question 1)
  - Understand that the size of a sample is not related to making a causation conclusion (similar as question 2)
  - Know that a correlational study design does not imply causation.

Question 4 refers to the following situation:

On the first day of her statistics class, Dr. Smith gave students a pretest to determine their statistical knowledge. At the end of the course, she gave students the exact same test. Dr. Smith constructed the scatterplot (below on the left) between students' pretest and posttest scores. The solid point in the upper right corner of the scatterplot represents the pretest and posttest scores for John. It turns out that John's pretest score was actually 5 (not 100 as previously recorded), and his posttest score was 100. John's scores were corrected and a new scatterplot was constructed (below on the right).

Scatterplot with John's Incorrect Scores Scatterplot with John's Correct Scores



Area of reasoning: (2) Reasoning about characteristics of distribution
Learning goal: Able to reason about the effect of moving an influential point in a scatterplot to a new location on the correlation coefficient

4. How would you expect the strength of the correlation between the pretest and posttest scores for the new scatterplot with John's actual scores (above, right) to compare to the strength of the relationship for the original scatterplot (above, left)?

a. The new correlation would be weaker than the original correlation.

b. The new correlation would be stronger than the original correlation.

c. The new correlation would have the same strength as the original correlation.

Behavior

- To answer this question correctly, students need to
    - Know how to interpret a scatterplot
    - Know how to estimate the strength of the correlation by inspecting a scatterplot.

Area of reasoning: (2) Reasoning about characteristics of distribution
Learning goal: Able to reason about factors that affect the mean and median

5. In 2011, it was reported that the mean home price in the Hamptons (New York) increased by 20% within a single year, while the median home price decreased by 2% during that same year. Which of the following is the best explanation for this occurrence?

a. The price of most homes in the Hamptons decreased and more homes were sold in the Hamptons that year.
b. The reporters made an error in presenting the results; if the mean home price increases, the median home price must also increase.
c. Most of the homes in the Hamptons decreased in price and a small number of homes had large increases in price.

Behavior

- To answer this question correctly, students need to
    - Know that the median is more resistant to outliers than the mean.

- Know the relationship between mean and median without a visual representation.
- Use the relationship between mean and median in the context of an increase and decrease.

Question 6 refers to the following situation:

A researcher investigated the impact of a particular herbicide on the enzyme level of carbonyl reductase in fish. In the study, 60 farm-raised fish were randomly assigned to the treatment group (in which they were exposed to the herbicide) or to the control group (in which they were *not* exposed to the herbicide). There were 30 fish assigned to each group. After the study, the data were analyzed, and the results of that analysis are reported in the output below.

Two Sample t-test

data: enzyme by exposed
t = -0.9141, *p*-value = 0.3644

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
-11.149231 4.158601

sample estimates:
mean in group NO mean in group YES
47.72625 51.22157

For question 6, indicate if the statement is a valid or invalid inference that can be made from the study results.

Area of reasoning: (3) Reasoning about interpretations of statistical inferences
Learning goal: Able to reason that a large *p*-value does not provide significant evidence of an effect.

6. Based on the results of this study, the researchers should not conclude that the herbicide has an effect on the enzyme levels of farm-raised fish.

a. Valid
b. Invalid

Behavior
- To answer this question correctly, students need to
  - Know that a high p-value does not provide significant evidence of an effect.

Area of reasoning: (2) Reasoning about characteristics of distribution
Learning goal: Able to reason about the meaning of variability in the context of repeated measurements and in a context where small variability is desired.

7. Jean is considering two different routes for commuting to school. She sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the minutes of travel time for 5 days of travel on each route.

Route #1: 16, 11, 23, 7, 18 (Mean = 15, Standard Deviation = 6.20)

Route #2: 19, 15, 17, 16, 18 (Mean = 17, Standard Deviation = 1.58)

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early. Based on the data gathered and Jean's preferences, which route would you advise her to choose?

a. Route #1 since on average, she got to school quicker.

b. Route #2 since her travel times were more consistent.

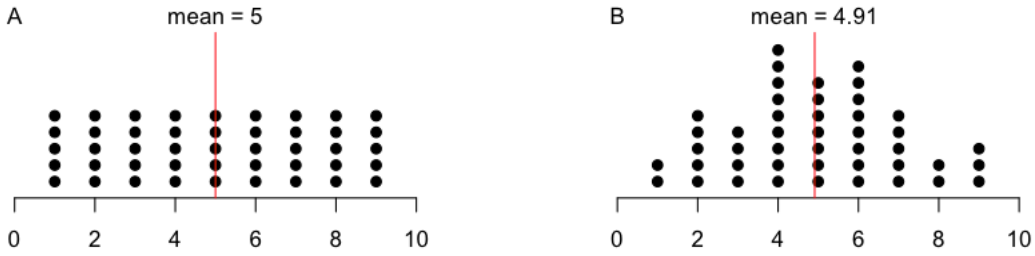
c. Jean can choose either route since the times are about the same.

Behavior

- To answer this question correctly, students need to
    - Know what the standard deviation represents
    - Know that a smaller standard deviation means the scores are more consistent.
    - Make connections between the mean and standard deviation.

Area of reasoning: (2) Reasoning about characteristics of distribution
Learning goal: Able to reason that given two distributions that have the same range, the one with less mass in the center has the larger standard deviation (tests for misconception that a uniform distribution has less "variability" than a non-uniform distribution)

8. Indicate which distribution has the larger standard deviation.



a. A has a larger standard deviation than B.

b. B has a larger standard deviation than A.

c. Both distributions have the same standard deviation.

Behavior

- To answer this question correctly, students need to
    - Know that the standard deviation measures the average distance from the mean.
    - Know how to identify variation on a dotplot (know that if there is more data closer to the mean and fewer far from the mean (graph B) than the standard

deviation will be smaller when compared to a uniform distribution.)
- Compare dotplots based on variation.

Question 9 refers to the following situation:

One hundred student-athletes attended a summer camp to train for a particular track race. All 100 student-athletes followed the same training program in preparation for an end-of-camp race. Fifty of the student-athletes were randomly assigned to additionally participate in a weight-training program along with their normal training (the training group). The other 50 student-athletes did not participate in the additional weight-training program (the non-training group). At the end of the summer camp, all 100 student-athletes ran the same race and their individual times (in seconds) were recorded.

The mean speed of the training group was 44 seconds, and the mean speed of the non-training group was 66 seconds.

Area of reasoning: (4) Reasoning about making inferences about group differences
Learning goal: Able to reason about how differences in variability affect strength of evidence against the null hypothesis of no difference

9. The standard deviation for the non-training group was 20 seconds. Consider the following possible values for the standard deviation of the training group. Which of these values would produce the strongest evidence of a difference between the two groups?

a. 10 seconds
b. 20 seconds
c. 30 seconds

Behavior
- To answer this question correctly, students need to
  - Know that differences in variability affect strength of evidence against the null hypothesis.
  - Know that if the mean is smaller but the variation is too large, then there might not be an evidence of a different between groups.

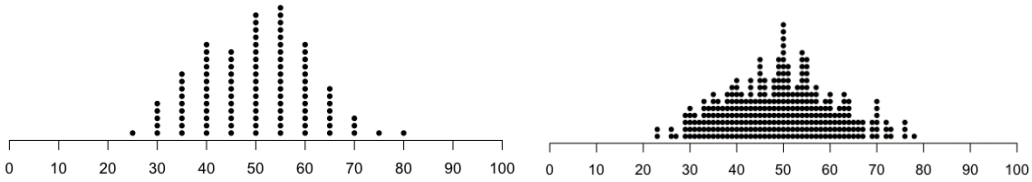
Question 10 refers to the following situation:

Imagine a candy company that manufactures a particular type of candy where 50% of the candies are red. The manufacturing process guarantees that candy pieces are randomly placed into bags. The candy company produces bags with 20 pieces of candy and bags with 100 pieces of candy.

Area of reasoning: (5) Reasoning about the relationship between sample size and sampling variability
Learning goal: Able to reason about differences between distributions of sample proportions for large and small sample sizes.

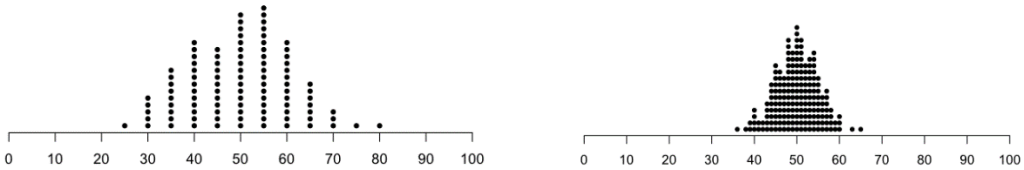
10. Which pair of distributions (below) most accurately represents the variability in the percentage of red candies in an individual bag that would be expected from many different bags of candy for the two different bag sizes?

a) 20-Piece Bags 100-Piece Bags



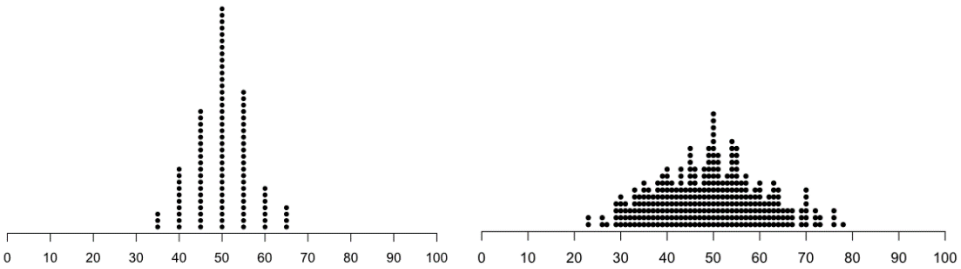
Percentage of Candies in a Bag that are Red Percentage of Candies in a Bag that are Red

b) 20-Piece Bags 100-Piece Bags



Percentage of Candies in a Bag that are Red Percentage of Candies in a Bag that are Red

c) 20-Piece Bags 100-Piece Bags



Percentage of Candies in a Bag that are Red Percentage of Candies in a Bag that are Red

Behavior

- To answer this question correctly, students need to
  - know how the standard error is affected by the sample size: (larger sample size → standard error decreases)

- Know how to identify a graph with less variability (smaller standard error).

Question 11 refers to the following situation:

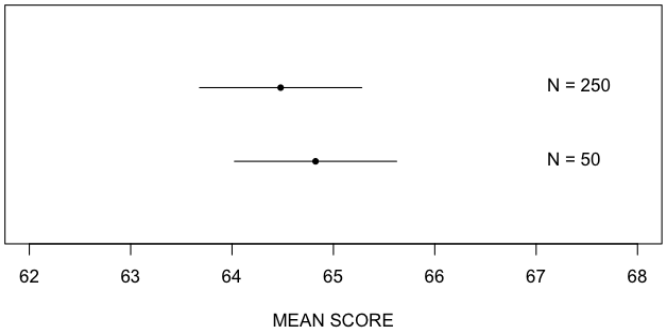
Question 11 asks you to think about factors that might affect the width of a confidence interval. A confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval.

Area of reasoning: (5) Reasoning about the relationship between sample size and sampling variability
Learning goal: Able to reason about how the width of a confidence interval is related to sample size.

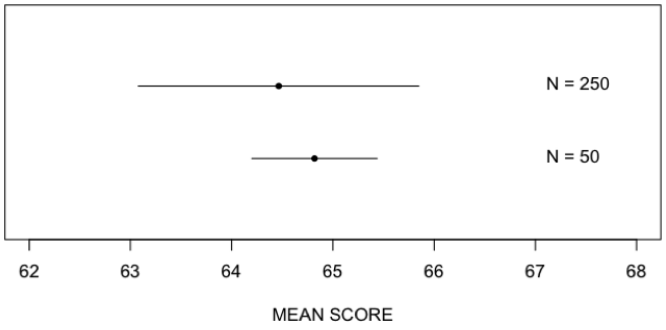
11. Imagine that two different random samples of test scores are drawn from a population of thousands of test scores. The first sample includes 250 test scores and the second sample includes 50 test scores. A 95% confidence interval for the population mean is constructed using each of the two samples.

Which set of confidence intervals (below) represents the two confidence intervals that would be constructed?

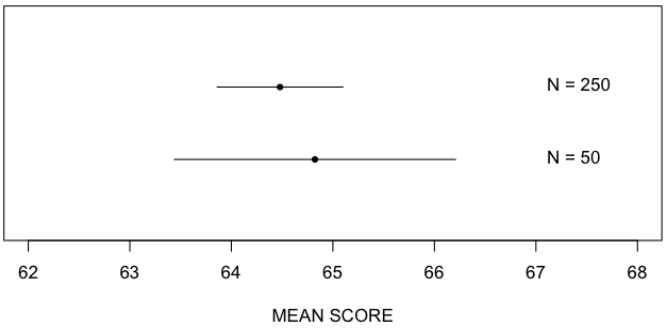
a.



b.



c.



Behavior

- To answer this question correctly, students need to
    - Know the relationship between sample size and variability.
    - Know how variability is related to the width of a confidence interval (width of a confidence interval decreases if the sample size increases)

Area of reasoning: (2) Reasoning about characteristics of distribution
Learning goal: Able to reason about the type of graphic representation that is needed to represent the shape, center, and variability of the distribution of a variable.

12. A teacher keeps track of the time it took her students to complete a particular exam (in minutes). These times are recorded in the table below.

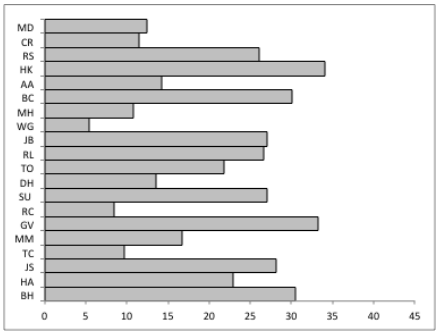
Student	Time
BH	30.5
HA	22.9
JS	28.1
TC	9.7
MM	16.7
GV	33.3
RC	8.5

Student	Time
SU	27.0
DH	13.6
TO	21.8
RL	26.7
JB	27.0
WG	5.4
MH	10.8

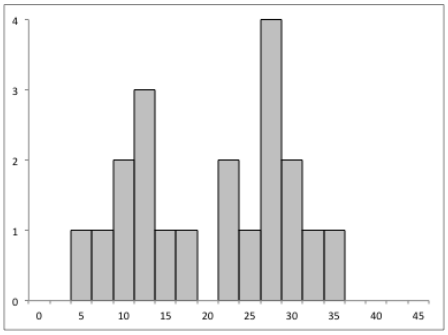
Student	Time
BC	30.1
AA	14.3
HK	34.1
RS	26.1
CR	11.5
MD	12.5

Each of the graphs below presents a valid representation of the time taken to complete the exam. Which of the graphs is the most appropriate display of the distribution of the times, in that the graph allows the teacher to describe the shape, center, and variability of the completion times?

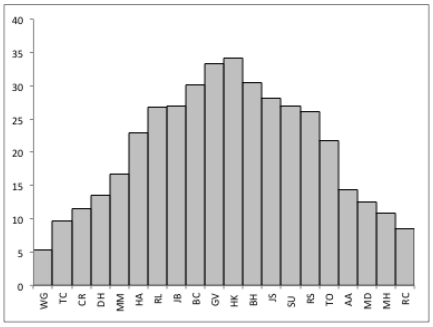
A.



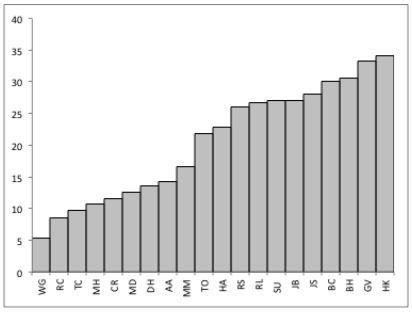
B.



C.



D.



Behavior
- To answer this question correctly, students need to
    - Know what type of visual representation allows for a better understanding of the shape, location, and variation of a distribution of the completion times variable.

Know that even if a distribution is normally distributed, it might not be the best visual representation.

Question 13 refers to the following situation:

A group of researchers investigated the impact of every-other-day fasting on weight loss in humans. After 22 days of every-other-day fasting, the researchers measured the amount of weight that was lost by each of 16 human subjects. These data were used to construct a 95% confidence interval for the mean weight loss for the population of people who followed the every-other-day fasting diet for 22 days. The computed interval was 4.1 ± 0.7 pounds. The interpretation of the interval given by the researchers was,

> *With 95% confidence, we can infer that a randomly selected person from the population who follows the every-other-day fasting diet for 22 days would lose between 3.4 and 4.8 pounds.*

Area of reasoning: (3) Reasoning about interpretations of statistical inferences
Learning goal: Able to reason about a misinterpretation of a confidence level (using it to make a prediction for a single case)
13. Why is the researchers' interpretation of the confidence interval incorrect?

a. Because anything could happen for the people on the diet; some could lose less than 3.4 or more than 4.8 pounds.
b. Because the confidence interval estimates the weight loss for the sample, and not the weight loss for the population.
c. Because the confidence interval estimates the population mean weight loss and not an individual's weight loss.
d. Because 95% of people who follow the diet will be in this range, losing between 3.4 and 4.8 pounds.

Behavior
- To answer this question correctly, students need to
    - Know that when interpreting a confidence interval the inferences are about a parameter (measure from a population) and not about individual subjects.
    - Know that a confidence interval gives plausible/likely values for a population parameter.

Area of reasoning: (3) Reasoning about interpretations of statistical inferences
Learning goal: Able to reason that a smaller p-value provides stronger evidence against the null hypothesis than a larger p-value.
14. Two medical researchers each perform the same experiment using two different samples from the same population. One study results in a p-value of 0.06, and the other study results in a *p*-value of 0.09. Which of the following statements is correct regarding the evidence against the

null hypothesis?

a. The *p*-value of 0.06 gives stronger evidence against the null hypothesis because it is smaller.
b. The *p*-value of 0.09 gives stronger evidence against the null hypothesis because it is larger.
c. It's impossible to tell which *p*-value provides stronger evidence against the null hypothesis, because they are both greater than 0.05.

Behavior
- To answer this question correctly, students need to
    - Know how a p-value can be used as evidence against the null.
    - Know that a smaller p-value gives more evidence against the null hypothesis (even if they are both greater than 0.05).
    - Know that the cut off level of 0.05 is not set in stone and the size of the p-value influences the conclusions that will be done.

Questions 15 to 20 refers to the following situation:
Yolanda was interested in whether offering people financial incentives can improve their performance playing video games. Yolanda designed a study to examine whether video game players are more likely to win a game when they receive a $5 incentive or when they simply receive verbal encouragement. Forty subjects were randomly assigned to one of two groups. The first group was told they would receive $5 if they won the game and the second group received verbal encouragement to "do your best" on the game. Yolanda collected the following data from her study:

	$5 Incentive	**Verbal Encouragement**	**Total**
Win	16	8	*24*
Lose	4	12	*16*
Total	*20*	*20*	*40*

Based on these data, it appears that the $5 incentive was more successful in improving performance than the verbal encouragement, because the observed difference in the proportion of players who won was

$$\frac{16}{20} - \frac{8}{20} = \frac{8}{20} = 0.40$$

In order to test whether this observed difference is only due to chance, Yolanda does the following:

- She gets 40 index cards. On 24 she writes, "win" and on 16 she writes, "lose."
    - She then shuffles the cards and randomly places the cards into two stacks of 20

cards each. One stack represents the participants assigned to the $5 incentive group and the other represents the participants assigned to the verbal encouragement group.

- She computes the difference in performance for these two hypothetical groups by subtracting the proportion of winning players in the "verbal encouragement" stack from the proportion of winning players from the "$5 incentive" stack.
- She records the computed difference on a plot

Yolanda repeats the previous three steps 100 times.

Area of reasoning: (4) Reasoning about making inferences about group differences
Learning goal: Able to reason about what the null model represents in a research study
15. What is the explanation for the process Yolanda followed?

a. This process allows her to determine the percentage of time the $5 incentive group would outperform the verbal encouragement group if the experiment were repeated many times.
b. This process allows her to determine how many times she needs to replicate the experiment for valid results.
c. This process allows her to see how different the two groups' performance would be if both types of incentive were equally effective.

Behavior
- To answer this question correctly, students need to
  - Know why someone would do a simulation
  - Know what the null model represents in the study (no difference/effect between the groups)
  - Know how the null model can be simulated.

Area of reasoning: (4) Reasoning about making inferences about group differences
Learning goal: Able to reason about what model should be used for the null hypothesis when comparing two groups
16. Yolanda simulated data under which of the following assumptions?

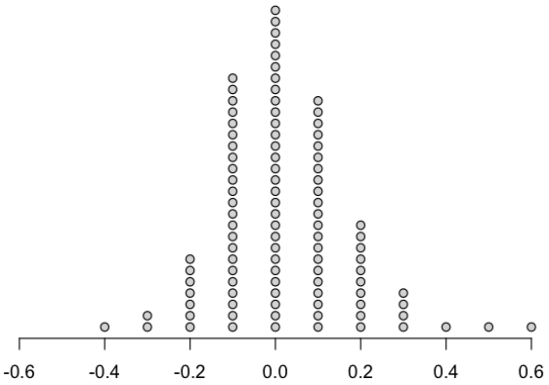
a. Verbal encouragement is more effective than a $5 incentive for improving performance.
b. The $5 incentive is more effective than verbal encouragement for improving performance.
c. The $5 incentive and verbal encouragement are equally effective at improving performance.

Behavior
- To answer this question correctly, students need to
  - Know that the data is being simulated under the null hypothesis of no difference.
  - Know what the null model represents in the study (the $5 incentive and

verbal encouragement are equally effective at improving performance)

Use the following for question 17:

Below is a plot of the simulated differences in proportion of wins that Yolanda generated from her 100 trials. Based on this plot, the one-sided *p*-value is 0.03.



(Proportion of Wins in $5 Incentive stack) – (Proportion of Wins in Verbal Encouragement stack)

Area of reasoning: (4) Reasoning about making inferences about group differences
Learning goal: Able to reason about a conclusion based on a statistically significant *p*-value in the context of a research study that compares two groups

17. Which of the following conclusions about the effectiveness of the $5 incentive is valid based on these simulation results?

a. The $5 incentive is more effective than verbal encouragement because the *p*-value is less than 0.05.
b. The $5 incentive is more effective than verbal encouragement because the distribution is centered at 0.
c. The $5 incentive is not more effective than verbal encouragement because the distribution is centered at 0.
d. The $5 incentive is not more effective than verbal encouragement because the *p*-value is less than 0.05.

Behavior

- To answer this question correctly, students need to
    - Know that a small p-value provides significant evidence of a difference between groups.
    - Know that the center of the distribution of simulated data is related to the null model hypothesis of no difference between groups.
    - Know that the center of the simulated distribution should not be used to make inferences about group difference.

For question 18, indicate whether the provided interpretation of the p-value is valid or invalid.

Area of reasoning: (3) Reasoning about interpretations of statistical inferences
Learning goal: Able to reason about an incorrect interpretation of a p-value (probability of a treatment being more effective).

18. The *p*-value is the probability that the $5 incentive group would win more often than the verbal encouragement group.

a. Valid

b. Invalid

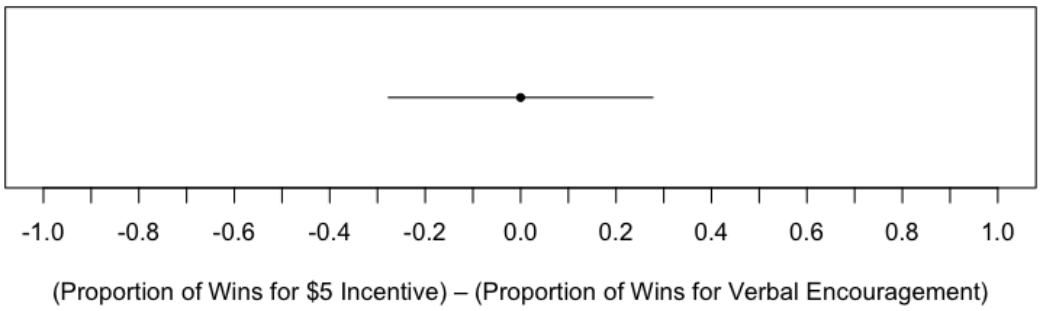
Behavior

- To answer this question correctly, students need to
    - Know how to interpret a p-value. Specifically, that the p-value is **not** the probability of a treatment being more effective.

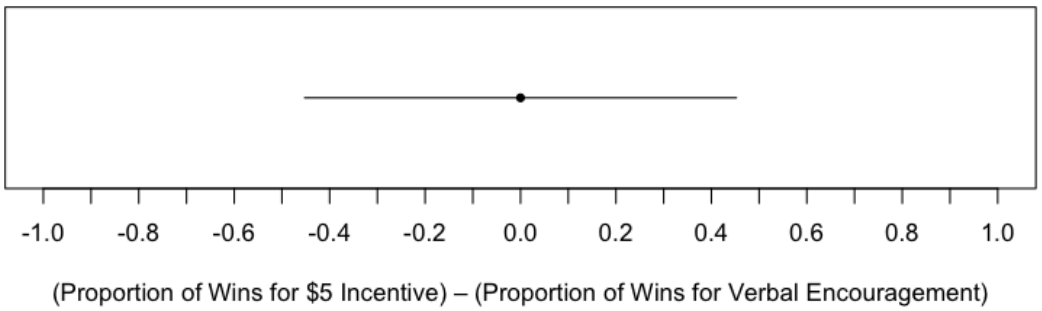
Area of reasoning: (4) Reasoning about making inferences about group differences
Learning goal: Able to reason about a confidence interval based on a statistically significant difference between groups.

19. In each graph below, a confidence interval is shown as a horizontal line. Which of the following graphs represents the 95% confidence interval for the true difference between the population proportions of players who receive a $5 incentive and players who receive verbal encouragement?

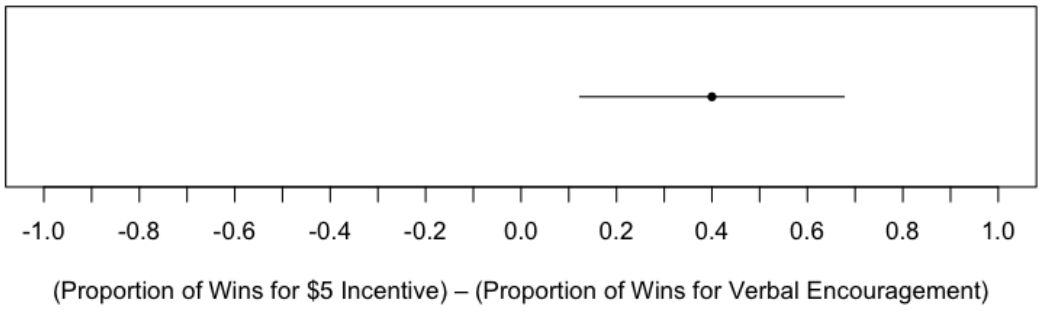
a.



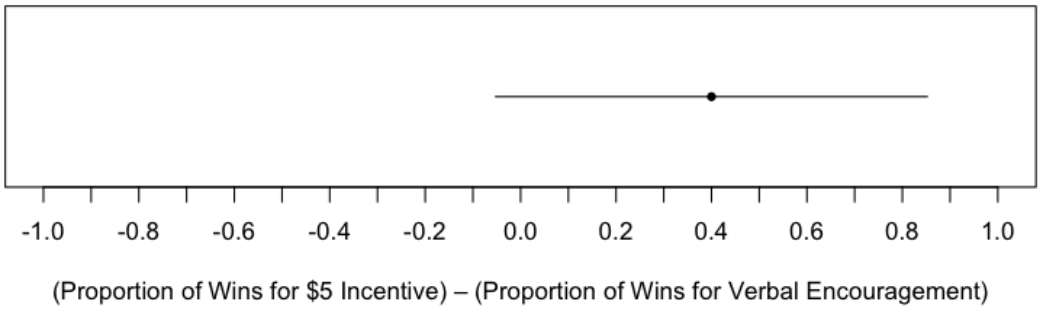
b.



c.



d.



Behavior
- To answer this question correctly, students need to
    - Make connections between the results from a hypothesis test and how to compute a confidence interval for the difference.

Students need to recognize that if a difference is found between groups, then the confidence interval for the difference in population proportion will not contain zero.

Area of reasoning: (5) Reasoning about the relationship between sample size and sampling variability

Learning goal: Able to reason about how increasing the sample size affects the p-value, all else being equal.

20. In Yolanda's experiment, there were 20 subjects randomly assigned to each group. Imagine a new study where 100 students were randomly assigned to each of the two groups. Assume that the observed difference in this new study was again 0.40 (i.e., that the proportion of wins for the $5 incentive group was 0.40 higher than the observed proportion of wins for the verbal encouragement group).

How would the *p*-value for this new study (100 per group) compare to the *p*-value for the original study (20 per group)?

a. It would be the same as the original *p*-value.

b. It would be smaller than the original *p*-value.

c. It would be larger than the original *p*-value.

Behavior
- To answer this question correctly, students need to
    - Know that if the sample size increases and the result from the study remains the same, then the p-value would get smaller.

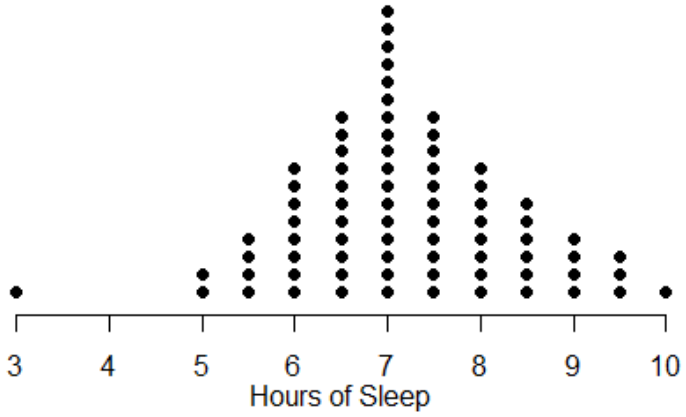
Know that the variability of a sampling distribution would decrease with a larger sample size.

Appendix C: Versions of the REALI assessment

C1 - Expert review version of the REALI assessment

1A

The following graph shows the distribution of hours slept the previous night by a group of college students.

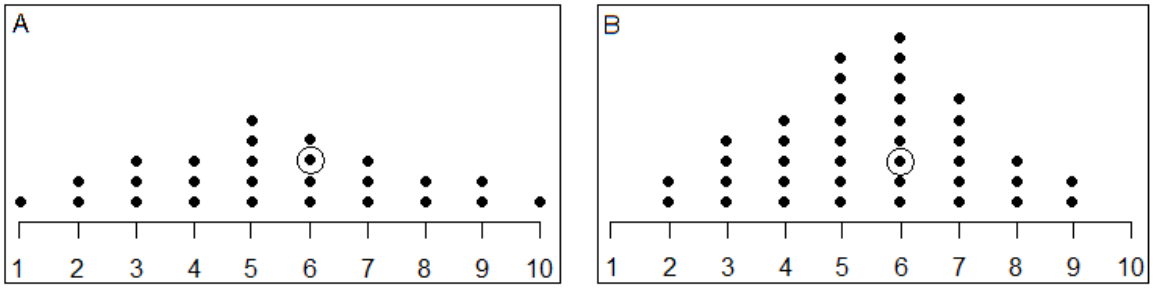


Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.

a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five.
b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
c. Many students seem to be getting 7 hours of sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.
d. The distribution of hours of sleep is somewhat normal, with an outlier at 3. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.

2A

Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights for 39 random samples of 3 pebbles each, with all mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.



a. No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
b. Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6.
c. Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B represents the average weight of 3 pebbles.

3A

A teacher keeps track of the time it took her students to complete a particular exam (in minutes). These times are recorded in the table below.

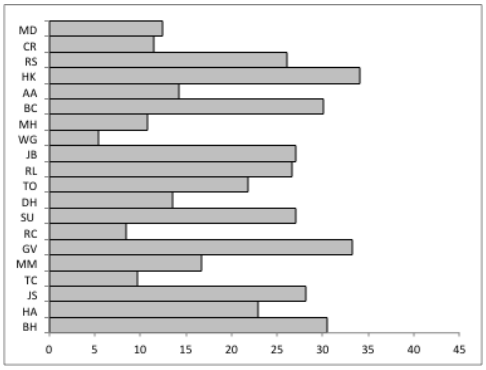
Student	Time
BH	30.5
HA	22.9
JS	28.1
TC	9.7
MM	16.7
GV	33.3
RC	8.5

Student	Time
SU	27.0
DH	13.6
TO	21.8
RL	26.7
JB	27.0
WG	5.4
MH	10.8

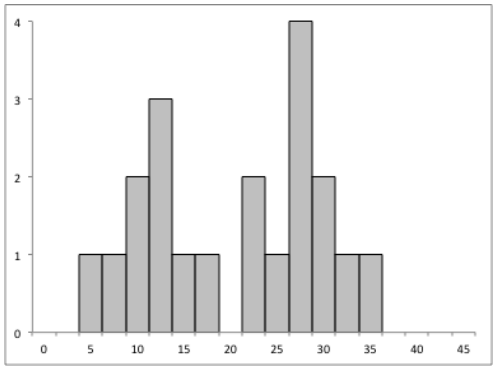
Student	Time
BC	30.1
AA	14.3
HK	34.1
RS	26.1
CR	11.5
MD	12.5

Each of the graphs below presents a valid representation of the time taken to complete the exam. Which of the graphs is the most appropriate display of the distribution of the times, in that the graph allows the teacher to describe the shape, center, and variability of the completion times?

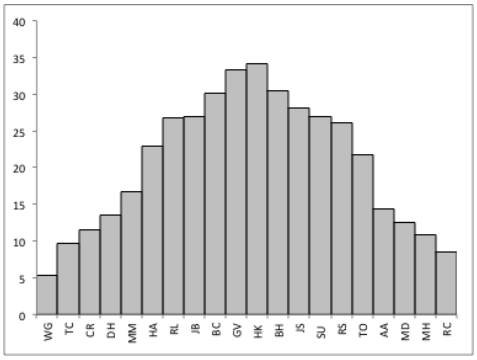
A.



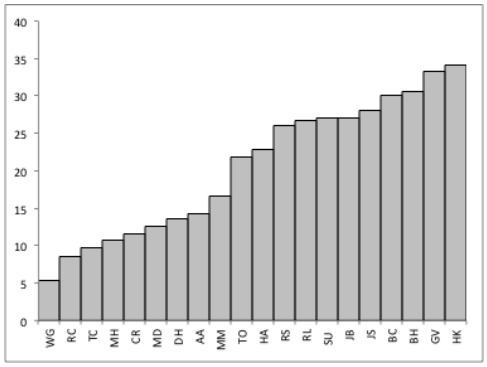
B.



C.

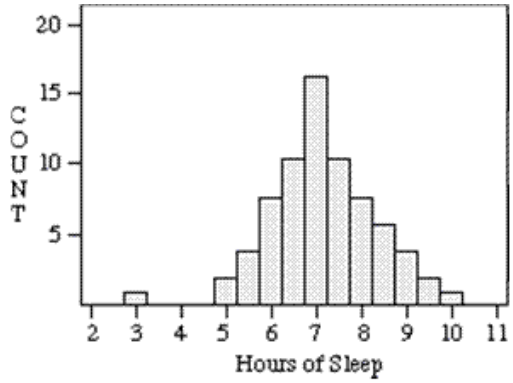


D.



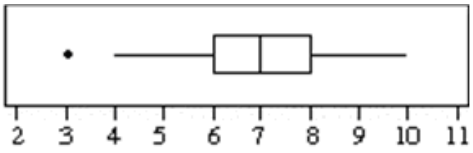
4A

The following graph shows a distribution of hours slept last night by a group of college students.

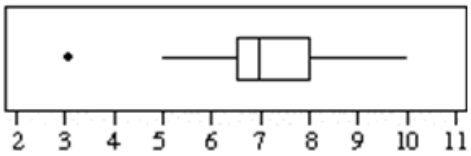


Which box plot seems to be graphing the same data as the histogram above?

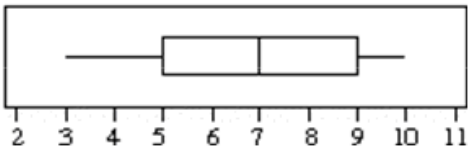
a)



b)



c)



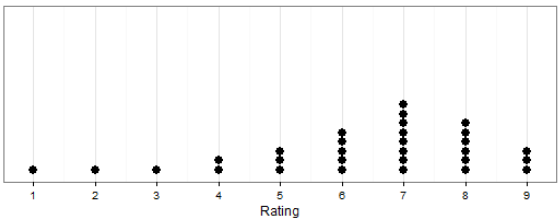
5A

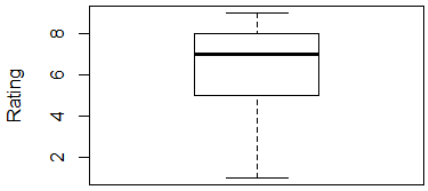
One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. After analyzing the answers from the students, the instructor interpret the data saying

> "**A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding."**

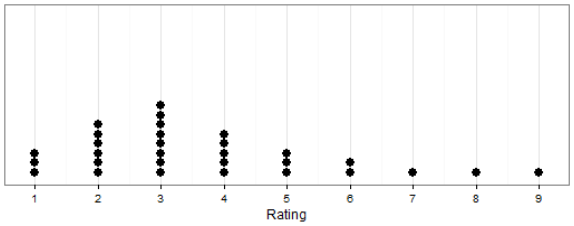
The instructor asked two of his students to create a graphical representation of the data, based on his interpretation above. Beth created a dotplot and Allan created a boxplot. Which dotplot/boxplot pair better aligns with the description given by the instructor?

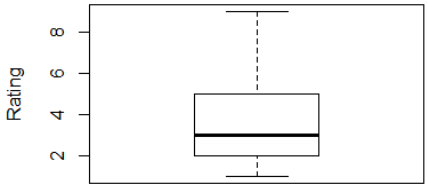
A



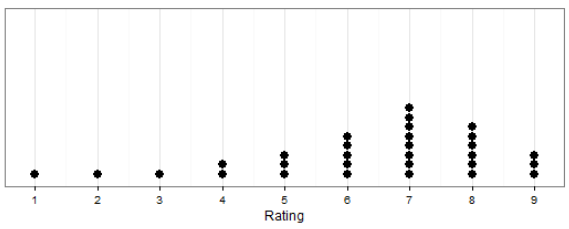


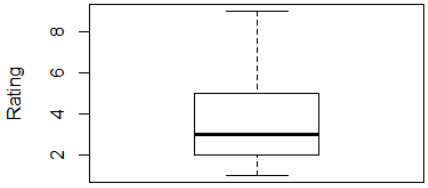
B



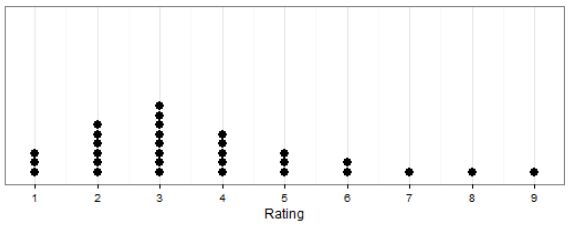


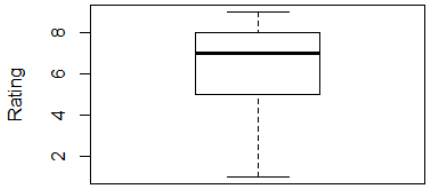
C





D





1G

Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 50 times and the other student flips a coin 100 times. Which student is more likely to get 48% to 52% of their coin flips heads up?

a. The student who flips the coin 50 times because the percent that are heads up is less likely to be exactly 50%.
b. The student who flips the coin 100 times because that student has more chances to get a coin flip that is heads up.
c. The student who flips the coin 100 times because the more flips that are made will increase the chance of approaching a result of 50% heads up.
d. Neither student is more likely because the flipping of the coin is random and therefore you cannot predict the outcome of the flips.

2G

According to the National Cancer Institute, the probability of a man in the United States developing prostate cancer at some point during his lifetime is 0.15. What does the statistic, 0.15, mean in the context of this report from the National Cancer Institute?

a. For all men living in the United States, approximately 15% will develop prostate cancer at some point in their lives.
b. If you randomly selected a male in the United States there is a 15% chance that he will develop prostate cancer at some point in his life.
c. In a random sample of 100 men in the United States, 15 men will develop prostate cancer.
d. Both *a* and *b* are correct.

3G

The Gopher 5 is a cash lotto game in Minnesota. To play, players pick five numbers from 1 to 47. Each number can only be used once. The numbers are listed in numerical order (not necessarily the order in which they were selected). A player wins the Gopher 5 Jackpot if all five numbers chosen by that player match the five winning numbers chosen randomly by a computer. Here are four sets of five numbers that players have chosen for the Gopher 5:

Set 1: 5 – 10 – 15 – 20 – 25
Set 2: 1 – 13 – 25 – 31 – 42
Set 3: 10 – 16 – 24 – 25 – 40
Set 4: 1 – 2 – 3 – 4 – 5

Which of these sets of numbers is less likely to win the Gopher 5?

A. Set 1
B. Set 2

C. Set 3
D. Set 4
E. None of the above.

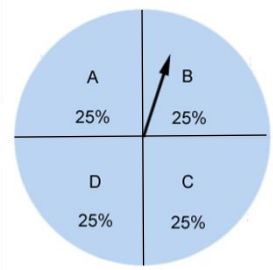
4G

A game company created a little plastic dog that can be tossed in the air. It can land either with all four feet on the ground, lying on its back, lying on its right side, or lying on its left side. However, the company does not know the probability of each of these outcomes. Which of the following methods is most appropriate to estimate the probability of each outcome?

A. Since there are four possible outcomes, assign a probability of 1/4 to each outcome.
B. Toss the plastic dog many times and see what percent of the time each outcome occurs.
C. Simulate the data using a model that has four equally likely outcomes.

5G

Consider a spinner shown below that has the letters from *A* to *D*.



Joan used the spinner 10 times and each time she wrote down the letter that the spinner landed on. When she looked at the results, she saw that the letter *B* showed up 5 times out of the 10 spins. Now she doubts the fairness of the spinner because it seems like she got too many *B*s.
A statistician wants to set up a probability model to examine how often the result of 5 B's out of 10 spins could happen with a fair spinner just by chance alone. Which of the following is the best probability model for the statistician to use?

A. The probability for each letter is the same - 1/4 for each letter.
B. The probability for letter B is 1/2 and the other three letters each have probability of 1/6.
C. The probability for letter B is 1/2 and the probabilities for the other letters sum to 1/2.

6G

Five faces of a fair die are painted black, and one face is painted white. The die is rolled six times. Which of the following results is more likely?

A. Black side up on five of the rolls; white side up on the other roll
B. Black side up on all six rolls
C. *a* and *b* are equally likely

1F

The following situation models the logic of a hypothesis test. An electrician tests whether or not an electrical circuit is good. The null hypothesis is that the circuit is good. The alternative hypothesis is that the circuit is not good. The electrician performs the test and decides to reject the null hypothesis. Which of the following statements is true?

a. The circuit is definitely not good and needs to be repaired.
b. The circuit is most likely not good, 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 might not be good.

2F

A researcher wants to answer the following research question: Do sleep apnea patients look more attractive after receiving a new treatment (continuous positive airway pressure therapy)? Twenty patients with sleep apnea were given the new treatment. Pictures of the patients were taken before the treatment as well as two months after the treatment. An independent rater was shown the before and after pictures for each patient in a random order and asked to determine which picture the patient looked more attractive in. Seventy percent (70%) of the patients were rated as more attractive in the after picture. Is there a need to conduct a hypothesis test to determine whether the treatment is effective, or could the researcher just use the sample statistic (70%) as evidence of the effectiveness?

a. The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%.
b. The researcher should conduct a hypothesis test because a hypothesis test is always appropriate.
c. The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.

3F

The Amherst H. Wilder Foundation conducts surveys with homeless people in Minnesota. In their 2012 survey of 4,465 homeless adults, one question asked was "In the last 30 nights, how many nights have you spent outside, in a vehicle or vacant building, or some other place not intended for housing?" One research question that the surveyors had was "Is there a difference between males and females with regards to the average number of nights spent in a place not intended for housing?" Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question?

a. There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
b. There is a difference between men and women in terms of the *number* of nights spent

in a place not intended for housing.

c. There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
d. There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.

4F

A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain?

a. A large *p*-value.
b. A small *p*-value.
c. The magnitude of a *p*-value has no impact on statistical significance.

5F

Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up different scents. An experiment was conducted with a dog in Japan who was trained to smell bowel cancer in stool samples. In a test, the dog was presented with five stool samples; one from a cancer patient and four from healthy people. The dog indicated which stool sample was from the cancer patient. This was repeated a total of 38 times. Out of the 38 tests, the dog correctly identified the cancer sample 37 times. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies cancer more than one fifth of the time. The p-value is less than .001. Assuming it was a well- designed study, use a significance level of .05 to make a decision.

a. Reject the null hypothesis and conclude that the dog correctly identifies cancer more than one fifth of the time.
b. There is enough statistical evidence to prove that the dog correctly identifies cancer more than one fifth of the time.
c. Do not reject the null hypothesis and conclude there is no evidence that the dog correctly identifies cancer more than one fifth of the time.

6F

One hundred student-athletes attended a summer camp to train for a particular track race. All 100 student-athletes followed the same training program in preparation for an end-of-camp race. Fifty of the student-athletes were randomly assigned to additionally participate in a weight-training program along with their normal training (the training group). The other 50 student-athletes did not participate in the additional weight-training program (the non-training group). At the end of the summer camp, all 100 student-athletes ran the same race and their individual times (in seconds) were recorded.

The mean speed of the training group was 44 seconds, and the mean speed of the non-training group was 66 seconds. The standard deviation for the non-training group was 20 seconds. Consider the following possible values for the standard deviation of the training group. Which of these values would produce the strongest evidence of a difference between the two groups?

A. 10 seconds
B. 20 seconds
C. 30 seconds

7F

Bob did a study where 40 subjects were randomly assigned to two groups (20 per group). He performed a study, analyzed the data, and found a *p*-value when testing if the mean difference between the groups was statistically significant. Imagine that Bob did a new study where 200 students were randomly assigned to two groups (100 per group). Assume that the observed mean difference for the new study is the same as the observed mean difference in original study. How would the *p*-value for new study (100 per group) compare to the *p*-value for original study (20 per group)?

A. It would be the same as the original *p*-value.
B. It would be smaller than the original *p*-value.
C. It would be larger than the original *p*-value.

8F

It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. 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 from the Department of Natural Resources took a random sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?

A. A random sample of a sample size of 100 with a sample mean of 12.1.
B. A random sample of a sample size of 36 with a sample mean of 11.5.
C. A random sample of a sample size of 100 with a sample mean of 11.5.
D. A random sample of a sample size of 36 with a sample mean of 12.1.

9F

A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes ($P = 0.287$). What does this mean?

A. The distribution of the GPAs for male and female scholarship athletes are identical except

for 28.7% of the athletes.

B. The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287.
C. There is a 28.7% chance that a pair of randomly chosen male and female scholarship athletes would have a significant difference assuming that there is no difference.
D. There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between male and female scholarship athletes as that observed in the sample assuming that there is no difference.

10F

Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached.
The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean?

A. The sample sizes were too small to detect a true difference between the coached and not-coached students.
B. The observed difference between coached and not-coached students could occur just by chance alone even if coaching really has no effect.
C. The increase in test scores makes no difference in getting into college since it is not statistically significant.
D. The study was badly designed because they did not have equal numbers of coached and not-coached students.

11F

A researcher is interested if there is a significant difference in the average number of hours watching television between males and females 8^{th} grade students in the US. After gathering and analyzing the data, the researcher found that the difference in the average number of hours between the two groups was 4.37 hours. The researcher conducted a test to verify if this difference in means was statistically significant and found a p-value of 0.001. What would be the correct conclusion the researcher needs to make?

A. The difference between groups in the average number of hours watching television did happen by chance because the p-value is so small.
B. The difference between groups in the average number of hours watching television is NOT statistically significant because the p-value is too small.
C. There IS strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there is NO difference.

D. There is NOT strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there IS a difference.

12F

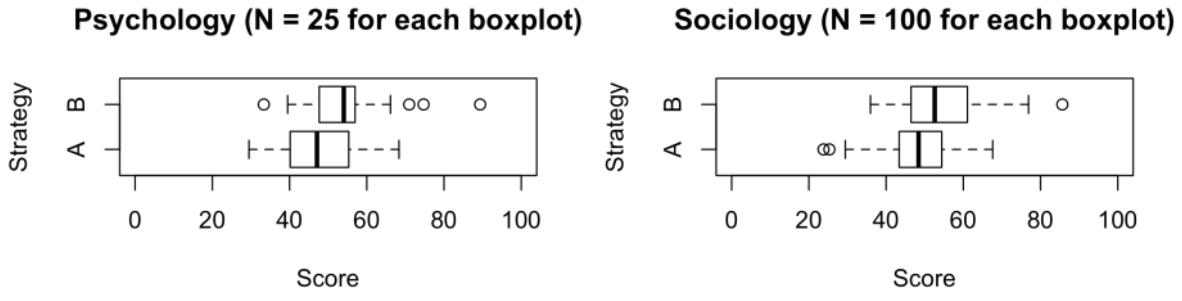
A research article reports the results of a new drug test. The drug is hypothesized to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article reports a *p*-value of 0.04 in the analysis section.
Which option below presents the correct interpretations of this p-value?

A. We conclude that the new drug is not effective because there is only a .04 probability that the drug is more effective than the current treatment.
B. We conclude that the new drug is effective because results like they found, or results even more favorable to the new drug, would only happen 4% of the time if the drug was not effective.
C. We conclude that the new drug is effective because there is only a 4% chance that it's not.
D. We conclude that the new drug is not effective because the difference in the proportion of macular degeneration patients with vision loss between the two treatments is only 0.04.

13F

Two experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100). The two different experiments were conducted with students who were enrolled in two different subject areas: psychology and sociology.

Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence against the claim, "neither strategy is better than the other"? Why?



A. Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment.
B. Psychology, because there are more outliers in strategy B from the Psychology

experiment, indicating that strategy B did not work well in that course.

C. Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment.
D. Sociology, because the sample size is larger in the Sociology experiment, which will produce a more accurate estimate of the difference between the two strategies.

14F

An engineer designs a new light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of 40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. A significant result for such a small difference would occur because:

A. The new design had more variability than the previous design.
B. The sample size for the new design is very large.
C. The mean of 1,200 for the previous design is large.

1C

Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quiz scores was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?

a. All of the individual scores are one point apart.
b. The difference between the highest and lowest score is 1 point.
c. The difference between the upper and lower quartile is 1 point.
d. A typical distance of a score from the mean is 1 point.

2C

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 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 above.

3C

Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following data for 5 days of travel on each route.

Country Route - 17, 15, 17, 16, 18

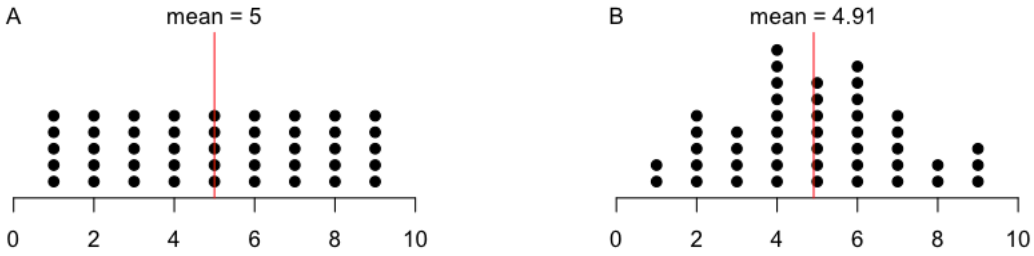
City Route - 18, 13, 20, 10, 16

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

A. The Country Route, because the times are consistently between 15 and 18 minutes.
B. The City Route, because she can get there in 10 minutes on a good day and the average time is less than for the Country Route.
C. Because the times on the two routes have so much overlap, neither route is better than the other. She might as well flip a coin.

4C

Indicate which distribution has the larger standard deviation.



A. A has a larger standard deviation than B.
B. B has a larger standard deviation than A.
C. Both distributions have the same standard deviation.

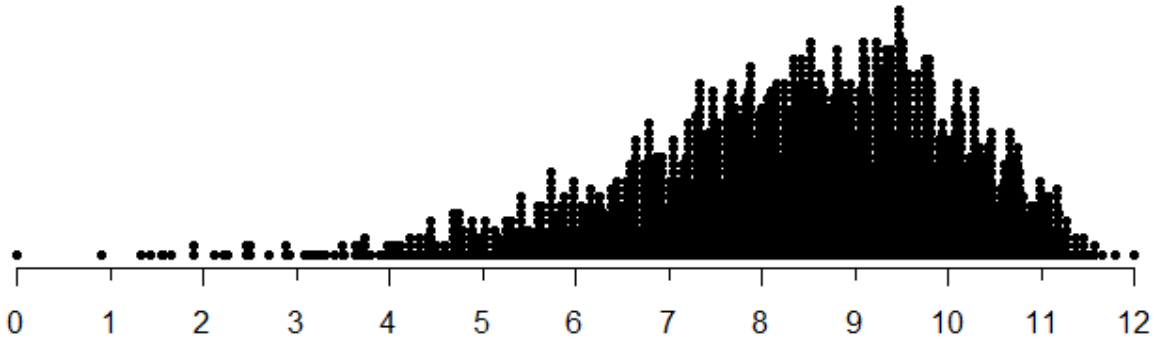
1B

According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the mean?

a. For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700.
For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700.
For all dog owners in this sample, about half were above $1,700 and about half were below $1,700.
For most owners, the first-year costs for owning a large-sized dog is $1,700.

2B

Which of the following intervals is MOST likely to include the mean of the distribution below?



a. 6 to 7
b. 8 to 9
c. 9 to 10
d. 10 to 11

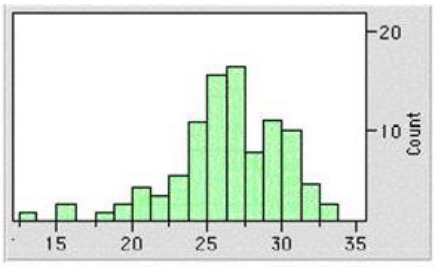
3B

In 2011, it was reported that the mean home price in the Hamptons (New York) increased by 20% within a single year, while the median home price decreased by 2% during that same year. Which of the following is the best explanation for this occurrence?

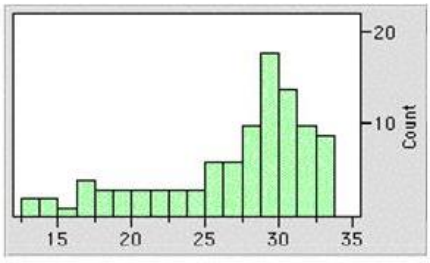
A. The price of most homes in the Hamptons decreased and more homes were sold in the Hamptons that year.
B. The reporters made an error in presenting the results; if the mean home price increases, the median home price must also increase.
C. Most of the homes in the Hamptons decreased in price and a small number of homes had large increases in price.

4B

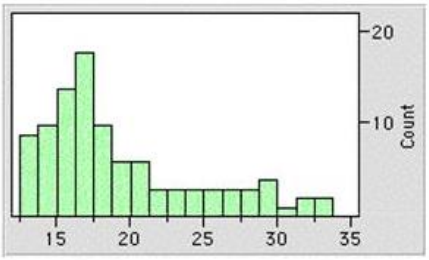
A study examined the length of a certain species of fish from one lake. The plan was to take a random sample of 100 fish and examine the results. The mean length was 26.8mm, the median was 29.4mm, and the standard deviation was 5.0mm. Which of the following histograms is most likely to be the one for these data?

A 

B



C



6B

The school committee of a small town wanted to determine the average number of children per household in their town. They divided the total number of children in the town by 50, the total number of households. Which of the following statements must be true if the average children per household is 2.2?

A. Half the households in the town have more than 2 children.
B. More households in the town have 3 children than have 2 children.
C. There are a total of 110 children in the town.
D. There are 2.2 children in the town for every adult.
E. The most common number of children in a household is 2.

1D

The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population.

a. The population is all American adults in 2013. The sample is the 21% of American adults that have had an email or social networking account compromised.
b. The population is the 1,002 American adults surveyed. The sample is all American adults in 2013.
c. The population is all American adults in 2013. The sample is the 1,002 American adults surveyed.

2D

CNN conducted a quick vote poll with a random sample of 5,581 Americans on September 19, 1999 to determine "What proportion of Americans still think that the Miss America pageant is still relevant today?" For the sample, 1,192 people voted yes and 4,389 people voted no. Identify the statistic and parameter of interest.

a. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the 5,581 Americans who took part in the survey.
b. The statistic is the 5,581 Americans who took part in the survey and the parameter is all Americans.
c. The statistic is the proportion of all Americans who think the pageant is still relevant and the parameter is the sample proportion of people who voted yes (1192/5581= .214).
d. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the proportion of all Americans who think the pageant is still relevant.

3D

A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level?

a. Observational study
b. Randomized experiment
c. Survey

4D

A research study randomly assigned participants into two groups. One group was given Vitamin E to take daily. The other group received only a placebo pill. The research study followed the participants for eight years. After the eight years, the proportion of each group that developed a particular type of cancer was compared.
What is the primary reason that the study used random assignment?

A. To ensure that the groups are similar in all respects except for the level of Vitamin E
B. To ensure that a person doesn't know whether or not they are getting the placebo.
C. To ensure that the study participants are representative of the larger population.

5D

A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.
Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

A. Valid, because the sample size is large enough to represent the population.
B. Valid, because 67% is far enough above 50% to predict a majority vote.
C. Invalid, because the sample is too small given the size of the population.
D. Invalid, because the sample may not be representative of the population.

6D

Researchers conducted a survey of 1,000 randomly selected adults in the United States and found a strong, positive, statistically significant correlation between income and the number of containers the adults reported recycling in a typical week.

Can the researchers conclude that higher income causes more recycling among U.S. adults? Select the best answer from the following options.

A. No, the sample size is too small to allow causation to be inferred.
B. No, the lack of random assignment does not allow causation to be inferred.
C. Yes, the statistically significant result allows causation to be inferred.
D. Yes, the sample was randomly selected, so causation can be inferred.

7D

A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?

A. This is a simple random sample. It will give an accurate estimate.
B. Because the sample is so small, it will not give an accurate estimate.
C. Because all fans had a chance to be asked, it will give an accurate estimate.
D. The sampling method is biased. It will not give an accurate estimate.

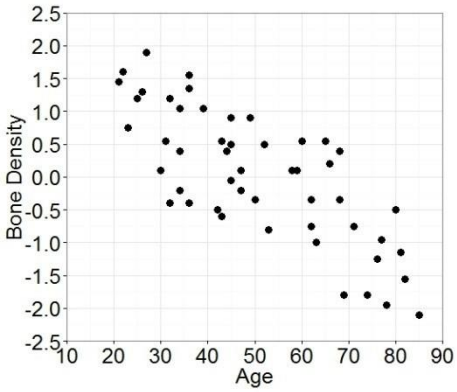
8D

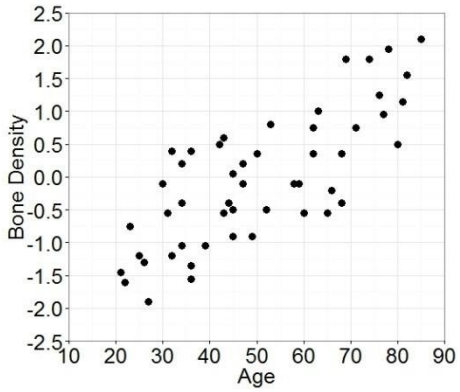
A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that 'prescription medications only' was the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?

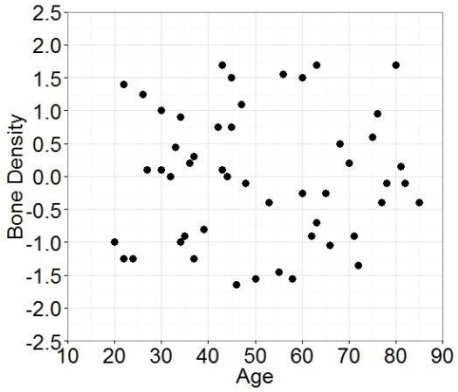
A. Yes, because medication patients lived longer.
B. No, because doctors chose the treatments.
C. Yes, because the study was a comparative experiment.
D. No, because the patients volunteered to be studied.

1H

Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?

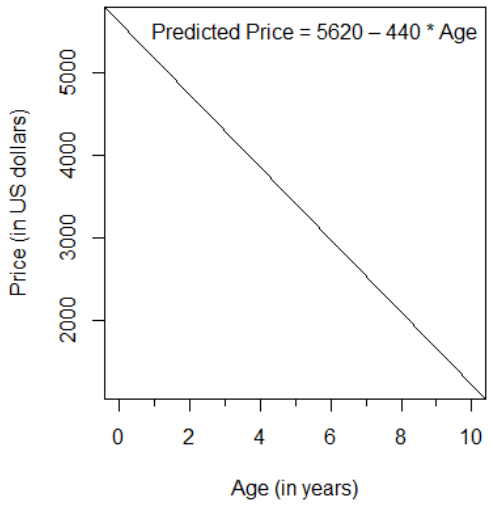






2H

A statistics student gathered data on a large numbers of cars of a particular model, from new cars to those that were up to 10 years old. Using the data on car ages (in years) and car prices (in US dollars) he found a linear relationship and produced the following regression equation and plot of the regression equation:



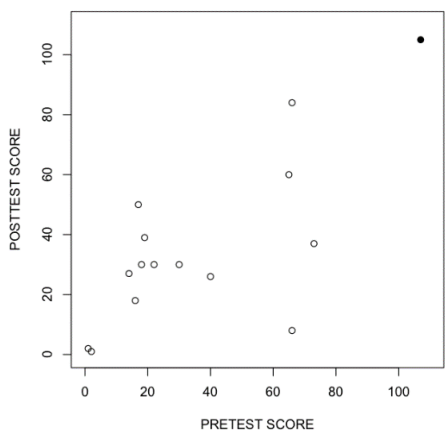
A friend asked him to use regression to predict the price of a 5 year-old model of this car. Which of the following methods can be used to provide an estimate?

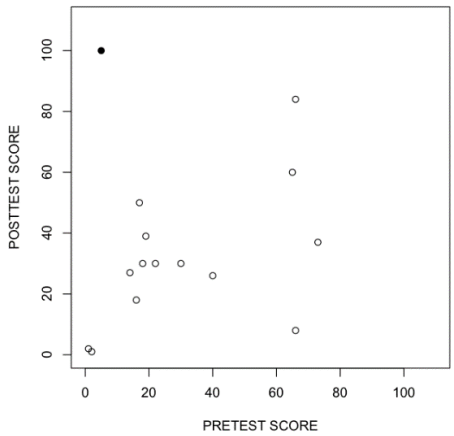
a. Locate the point on the line that corresponds to an age of 5 and read off the corresponding value on the y axis.
b. Substitute an age of 5 in the equation and solve for "Predicted Price".
c. Both of these methods are correct.
d. Neither of these methods is correct.

3H

On the first day of her statistics class, Dr. Smith gave students a pretest to determine their statistical knowledge. At the end of the course, she gave students the exact same test. Dr. Smith constructed the scatterplot (below on the left) between students' pretest and posttest scores. The solid point in the upper right corner of the scatterplot represents the pretest and posttest scores for John. It turns out that John's pretest score was actually 5 (not 100 as previously recorded), and his posttest score was 100. John's scores were corrected and a new scatterplot was constructed (below on the right).

Scatterplot with John's Incorrect Scores Scatterplot with John's Correct Scores



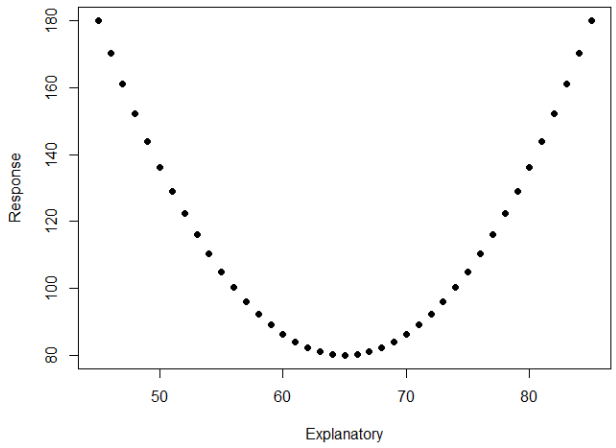


How would you expect the strength of the correlation between the pretest and posttest scores for the new scatterplot with John's actual scores (above, right) to compare to the strength of the relationship for the original scatterplot (above, left)?

A. The new correlation would be weaker than the original correlation.
B. The new correlation would be stronger than the original correlation.
C. The new correlation would have the same strength as the original correlation.

4H

The correlation between two variables is zero and the scatterplot for these variables is presented below.



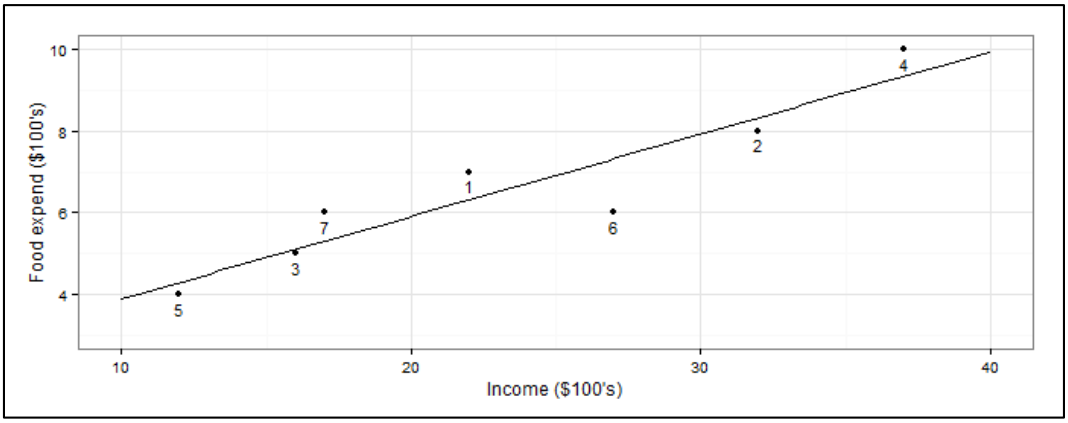
Based on the correlation value and the plot above, how would you interpret the relationship between the explanatory variable and the response variable?

A. There is no relationship between the explanatory variable and the response variable.

B. Correlation = 0 indicates a perfect relationship between an explanatory variable and a response variable.
C. Correlation is not appropriate to quantify the strength of the relationship between these variables.

5H

A random sample of 7 households was obtained, and information on their income and food expenditures for this year was collected. Below is a scatterplot of these data with the regression line superimposed:



The regression equation for these data is: ***predicted expend = 1.869 + 0.202 (income)***
After further analysis, the researcher found out that Point 4 was data from another year and decided to exclude it. What would be the regression equation after excluding Point 4?

A. predicted expend = 1.869 + 0.234 (income)
B. predicted expend = 1.869 + 0.202 (income)
C. predicted expend = 2.550 + 0.164 (income)
D. predicted expend = 1.821 + 0.234 (income)

6H

Past data was collected from students who take a certain statistics class, where test scores ranged from 0-100. The regression line relating the final exam score and the midterm exam score is:

Predicted final exam = 50 + 0.05(midterm)

What is a correct interpretation of the ***slope***?

A. A student who scored 0 on the final exam would be predicted to score 50 on the midterm exam.
B. As the score on the final increases by 1 point, it is predicted that the score on the midterm exam will increase by 0.05 point.
C. A student who scored 0 on the midterm would be predicted to score 50 on the final exam.

D. As the score on the midterm increases by 1 point, it is predicted that the score on the final exam will increase by 0.05 point.

1E

The Pew Research Center surveyed 2,076 American adult cell phone users chosen at random in 2013. The sample percent of adult cell phone users who access the internet on their phones was 60%. The 95% confidence interval was 58% to 62%. What is this interval attempting to estimate?

a. The average number of American adult cell phone users who access the internet on their phones in 2013.
b. The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
c. The percent of all American adult cell phone users who access the internet on their phones in 2013.
d. For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.

2E

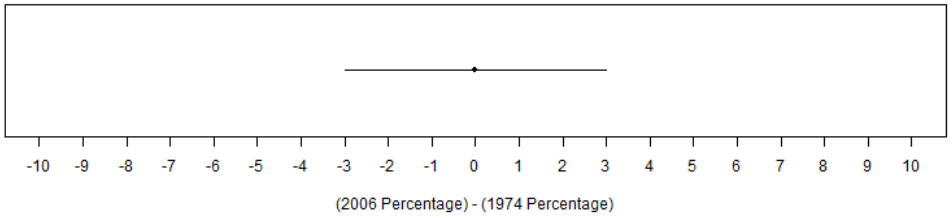
In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

a. We know that 37% of veterans in the *sample* have been divorced at least once.
b. We know that 37% of veterans in the *population* have been divorced at least once.
c. We can say with 95% confidence that 37% of veterans in the *sample* have been divorced at least once.
d. We can say with 95% confidence that 37% of veterans in the *population* have been divorced at least once.

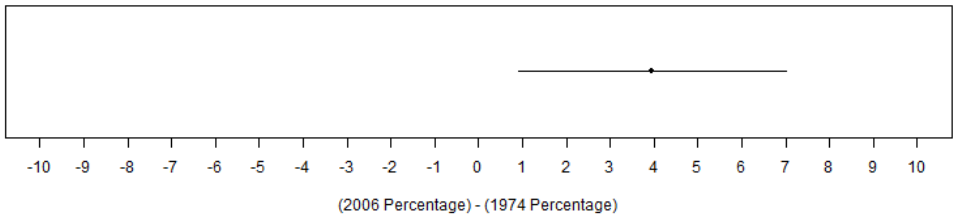
3E

A citizens' survey reported that in 1974, 64.5% of adults in a given state in the U.S. favored capital punishment, while in 2006 this percentage was 68.5%. To see if support for capital punishment has increased, a p-value for a test for difference in proportions is .011. Which of the following graphs represents a plausible 95% confidence interval for the difference in proportion of people who favored capital punishment between 1974 and 2006?

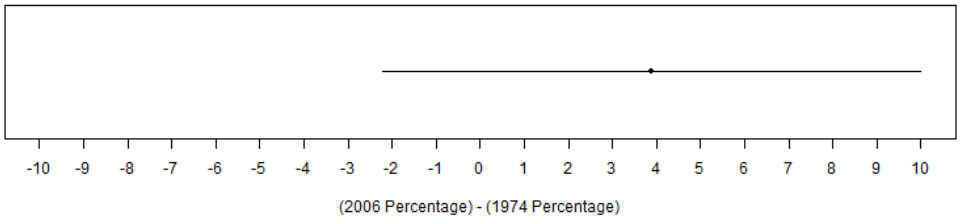
A



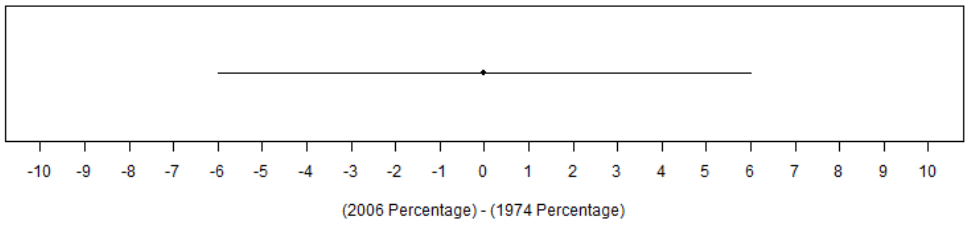
B



C



D



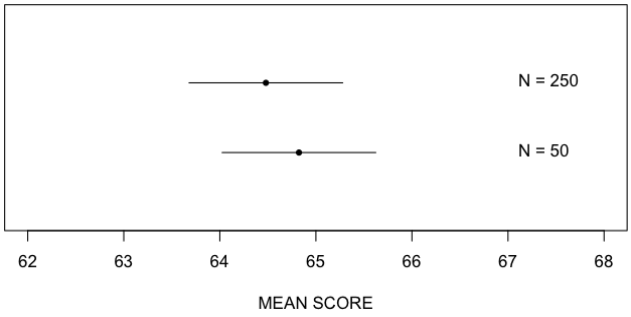
4E

Think about factors that might affect the width of a confidence interval. A confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval.

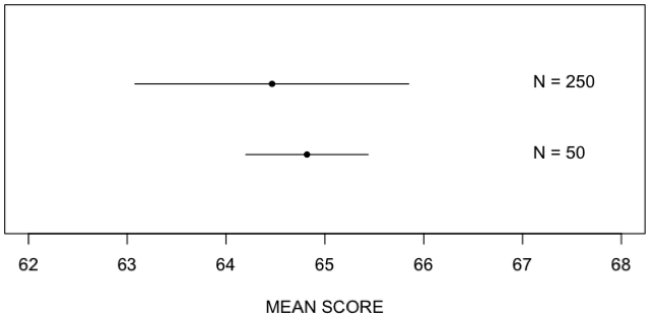
Imagine that two different random samples of test scores are drawn from a population of thousands of test scores. The first sample includes 250 test scores and the second sample includes 50 test scores. A 95% confidence interval for the population mean is constructed using each of the two samples.

Which set of confidence intervals (below) represents the two confidence intervals that would be constructed?

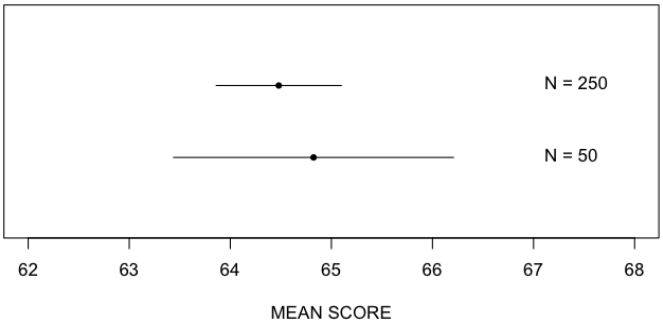
A.



B.



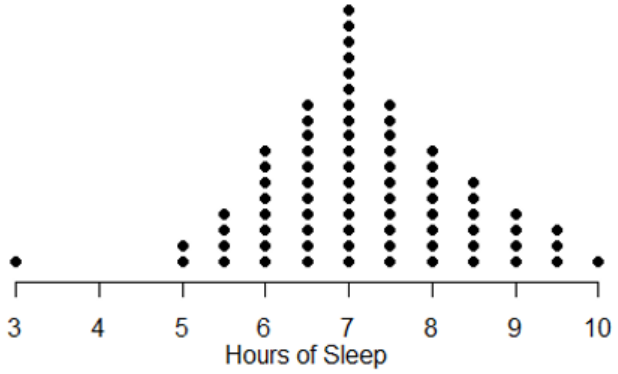
C.



C2 - Pilot study version of REALI assessment

1A

1. A high school teacher is concerned with whether her students are getting enough sleep. She asked each one of her students to report the number of hours slept the previous night. A plot of the results is given below.

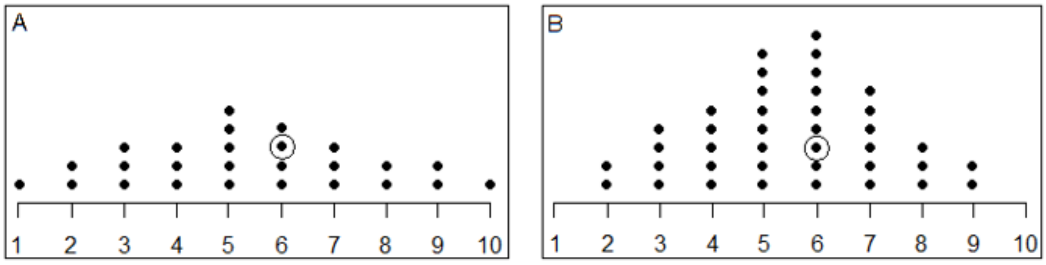


She would like to report her findings to the school principal using appropriate statistical language to describe and interpret the distribution in context. Select the **most appropriate** statement that she could use to summarize the results.

a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five. The standard deviation is about 2 hours.
b. The distribution of hours of sleep is a normal curve and centered at 7. There is a gap between three and five. The student who slept 3 hours may be an outlier.
c. The distribution of hours of sleep is somewhat bell-shaped. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.

2A

2. Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights for 39 random samples of 3 pebbles each, with all mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in Figure A and the dot circled in Figure B? Please select the best answer from the list below.



a. No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
b. Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots

with a weight of 6.

c. Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B is the mean weight of 3 pebbles.

3A

3. A teacher keeps track of the time it took her students to complete a particular exam (in minutes). These times are recorded in the table below.

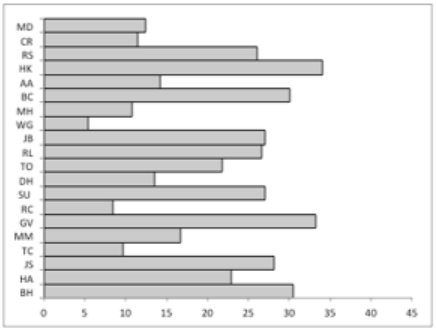
Student	Time
BH	30.5
HA	22.9
JS	28.1
TC	9.7
MM	16.7
GV	33.3
RC	8.5

Student	Time
SU	27.0
DH	13.6
TO	21.8
RL	26.7
JB	27.0
WG	5.4
MH	10.8

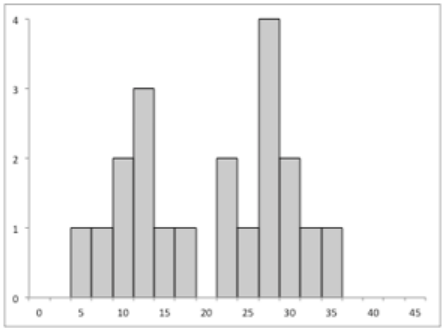
Student	Time
BC	30.1
AA	14.3
HK	34.1
RS	26.1
CR	11.5
MD	12.5

Each of the graphs below presents a valid representation of the time taken to complete the exam. Which of the graphs is the most appropriate display of the distribution of the times, in that the graph allows the teacher to describe the shape, center, and variability of the completion times?

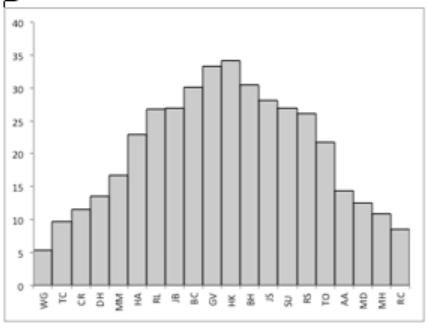
A.



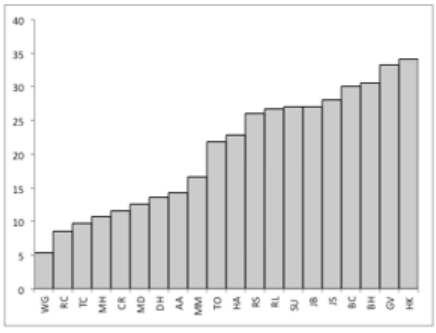
C



B

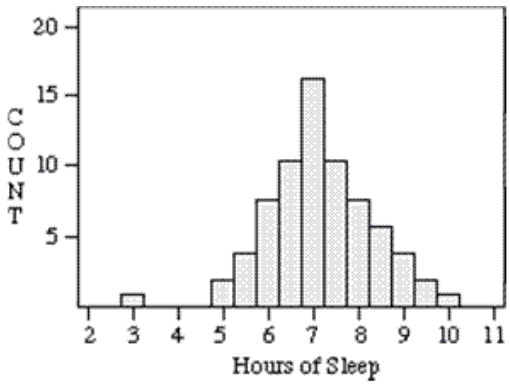


D.



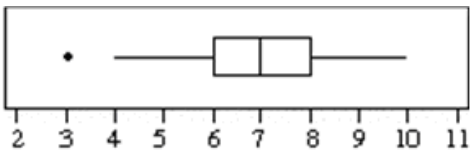
4A

4. The following graph shows a distribution of hours slept last night by a group of college students.

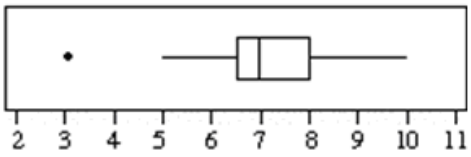


Which boxplot seems to be graphing the same data as the histogram above?

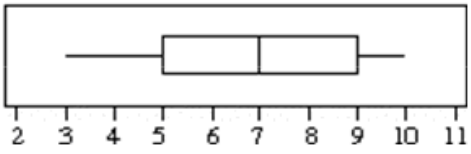
A



B



C



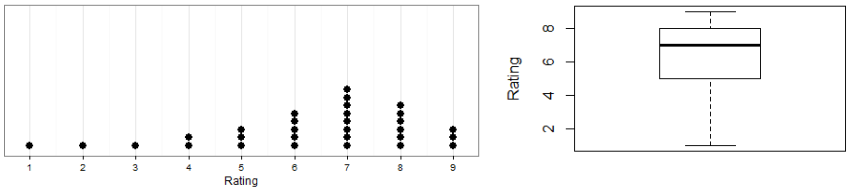
5A

5. An instructor gave his introductory statistics students a survey. One of the questions read, "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. After analyzing the answers from the students, the instructor interpreted the data as follows:

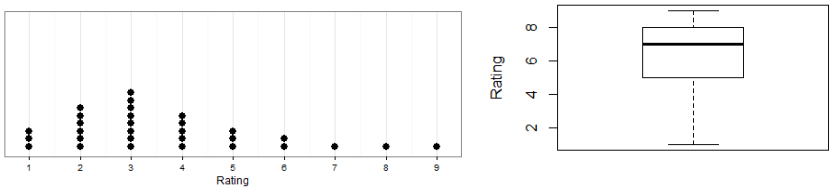
> "A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding."

The instructor asked two of his students to create a graphical representation of the data, based on his interpretation above. Beth created a dotplot and Allan created a boxplot. Which dotplot/boxplot pair best aligns with the description given by the instructor?

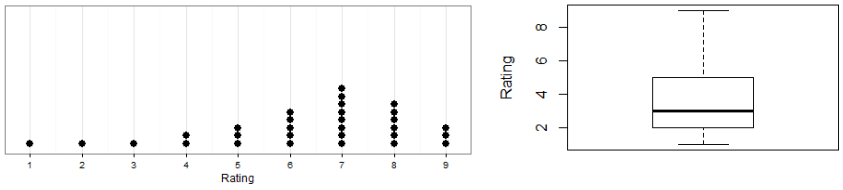
A



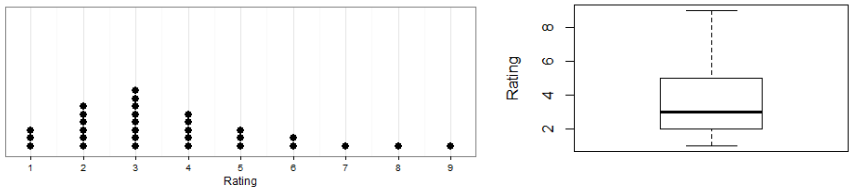
B



C

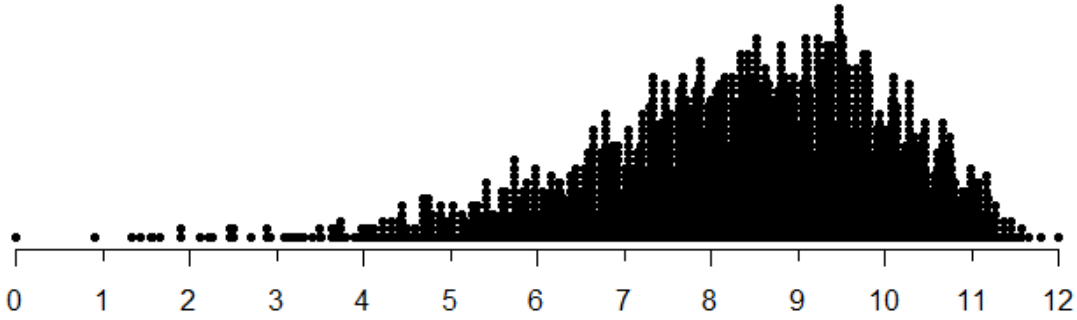


D



2B

6. Which of the following intervals is MOST likely to include the mean of the distribution below?



a. 6 to 7
b. 8 to 9
c. 9 to 10

1B

7. According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the average?

a. For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700.
b. For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700.
c. For all dog owners in this sample, about half were above $1,700 and about half were below $1,700.
d. For most owners, the first-year costs for owning a large-sized dog is $1,700.

3B

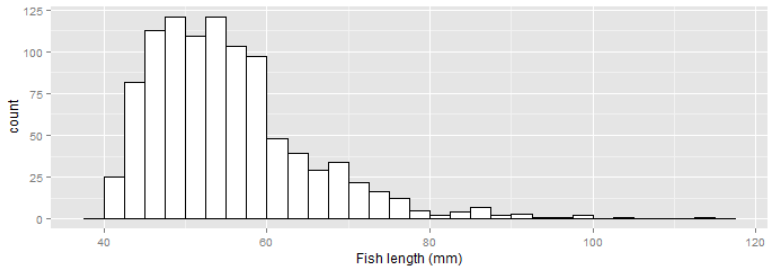
8. In 2011, it was reported that the mean home price in the Hamptons (New York) increased by 20% within a single year, while the median home price decreased by 2% during that same year. Which of the following is the best explanation for this occurrence?

a. The price of most homes in the Hamptons decreased and more homes were sold in the Hamptons that year.
b. The reporters made an error in presenting the results; if the mean home price increases, the median home price must also increase.
c. Most of the homes in the Hamptons decreased in price and a small number of homes had large increases in price.

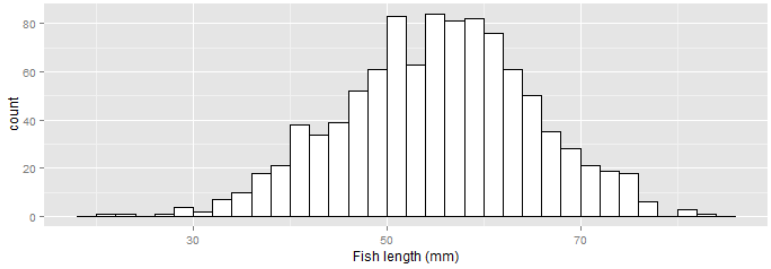
4B

9. A study examined the length of a certain species of fish from one lake. A random sample of 1000 fish had mean length of 52.2mm and median length of 57.4mm. Which of the following histograms is most likely to represent the data from this study?

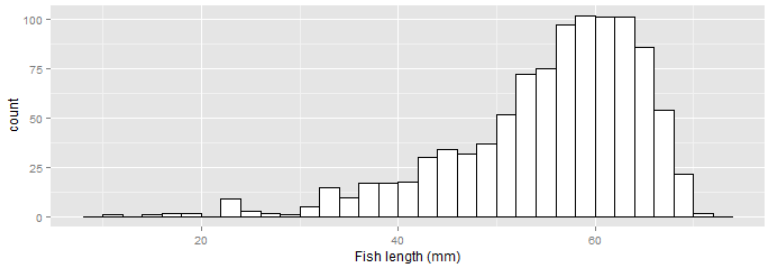
A



B



C



1C

10. Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quiz scores was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?

a. All of the individual scores are one point apart.
b. The difference between the highest and lowest score is 1 point.
c. The difference between the upper and lower quartile is 1 point.
d. A typical distance of a score from the mean is 1 point.

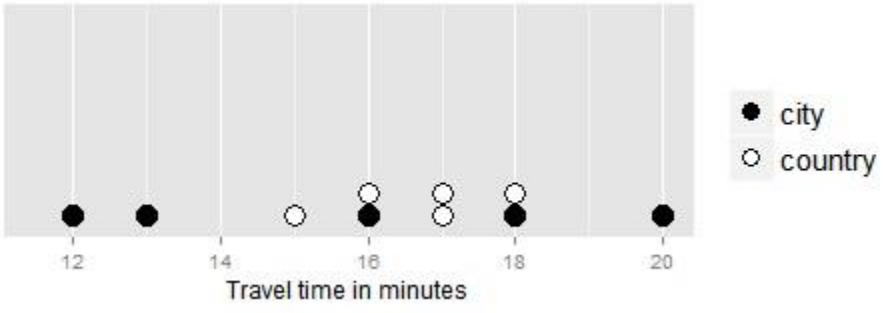
2C

11. A teacher gives a science test with 15 questions. For each question, 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 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 above.

3C

12. Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following commute times, in minutes, for 5 days of travel on each route.

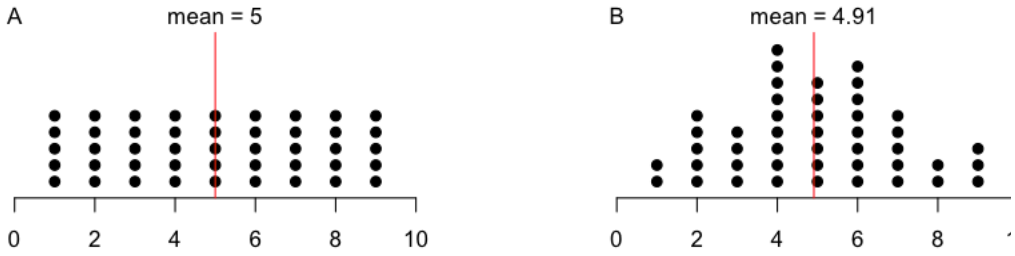


It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

a. The City Route
b. The Country Route
c. Neither route is better than the other.

4C

13. Indicate which distribution has the larger standard deviation.



a. A has a larger standard deviation than B.
b. B has a larger standard deviation than A.
c. Both distributions have the same standard deviation.

3D

14. A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level?

a. Observational study
b. Randomized experiment
c. Survey study

1D

15. The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results. Identify also the sample from that population.

a. The population is all American adults in 2013. The sample is the 21% of American adults that have had an email or social networking account compromised.
b. The population is the 1,002 American adults surveyed. The sample is all American adults in 2013.
c. The population is all American adults in 2013. The sample is the 1,002 American adults surveyed.

2D

16. CNN conducted a quick vote poll with a random sample of 5,581 Americans on September 19, 1999 to determine "What proportion of Americans still think that the Miss America pageant is relevant today?" For the sample, 1,192 people voted yes and 4,389 people voted no. Identify the statistic and parameter of interest.

a. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the 5,581 Americans who took part in the survey.
b. The statistic is the 5,581 Americans who took part in the survey and the parameter is all Americans.
c. The statistic is the proportion of all Americans who think the pageant is still relevant and the parameter is the sample proportion of people who voted yes (1192/5581= .214).
d. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the proportion of all Americans who think the pageant is still relevant.

5D

17. A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.
Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

a. Invalid, because the sample is only 1/50th of the size of the population.
b. Invalid, because not all viewers had the opportunity to respond to the poll.
c. Invalid, because the sample may not be representative of the population.

4D

18. A research study randomly assigned participants into two groups. One group was given Vitamin E to take daily (treatment 1). The other group received only a placebo pill (treatment 2). The research study followed the participants for eight years. After the eight years, the proportion of each group that developed a particular type of cancer was compared. What is the **primary** reason that the study used random assignment?

a. To ensure that the groups are comparable except for the treatment variable.
b. To ensure that a person doesn't know whether or not they are getting the placebo.
c. To ensure that the study participants are representative of the larger population.

6D

19. Researchers conducted a survey of 1,000 randomly selected adults in the United States and found a strong, positive, statistically significant correlation between income and the number of containers the adults reported recycling in a typical week.

Can the researchers conclude that higher income causes more recycling among U.S. adults? Select the best answer from the following options.

a. No, the sample size is too small to allow causation to be inferred.
b. No, the lack of random assignment does not allow causation to be inferred.
c. Yes, the statistically significant result allows causation to be inferred.
d. Yes, the sample was randomly selected, so causation can be inferred.

7D

20. A sportswriter wants to know the extent to which football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?

a. This is a simple random sample. It will give an accurate estimate.
b. Because the sample is so small, it will not give an accurate estimate.
c. Because all fans had a chance to be asked, it will give an accurate estimate.
d. The sampling method is biased. It will not give an accurate estimate.

8D

21. A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that 'prescription medications only' was the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?

a. Yes, because medication patients lived longer.
b. No, because doctors chose the treatments.
c. Yes, because the study was a comparative experiment.
d. No, because the patients volunteered to be studied.

2F

22. A researcher wants to answer the following research question: Do sleep apnea patients look more attractive after receiving a new treatment (continuous positive airway pressure therapy)? Twenty patients with sleep apnea were given the new treatment. Pictures of the patients were taken before the treatment as well as two months after the treatment. An independent rater was shown the before and after pictures for each patient in a random order and asked to determine which picture the patient looked more attractive in. Seventy percent (70%) of the patients were rated as more attractive in the after picture. Is there a need to conduct a hypothesis test to determine whether the treatment is effective, or could the researcher just use the sample statistic (70%) as evidence of the effectiveness?

a. The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%; therefore, the result did not happen by chance.
b. The researcher should conduct a hypothesis test because a hypothesis test is always appropriate.
c. The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.

4F

23. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain?

a. A large *p*-value.
b. A small *p*-value.
c. The magnitude of a *p*-value has no impact on statistical significance.

6F

24. One hundred student-athletes attended a summer camp to train for a particular track race. All 100 student-athletes followed the same training program in preparation for an end-of-camp race. Fifty of the student-athletes were randomly assigned to additionally participate in a weight-training program along with their normal training (the training group). The other 50 student-athletes did not participate in the additional weight-training program (the non-training group). At the end of the summer camp, all 100 student-athletes ran the same race and their individual times (in seconds) were recorded.

The mean speed of the training group was 44 seconds, and the mean speed of the non-training group was 66 seconds. The standard deviation for the non-training group was 20 seconds. Consider the following possible values for the standard deviation of the training group. Which of these values would produce the strongest evidence of a difference in means between the two groups?

a. 10 seconds
b. 20 seconds
c. 30 seconds

3F

25. The Amherst H. Wilder Foundation conducts surveys with homeless people in Minnesota. In their 2012 survey of 4,465 homeless adults, one question asked was "In the last 30 nights, how many nights have you spent outside, in a vehicle or vacant building, or some other place not intended for housing?" One research question that the surveyors asked was "Is there a difference between males and females with regards to the average number of nights spent in a place not intended for housing?"
Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question?

a. There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
b. There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
c. There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
d. There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.

5F

26. Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up different scents. An experiment was conducted with a dog in Japan who was trained to smell bowel cancer in stool samples. In a test, the dog was presented with five stool samples; one from a cancer patient and four from healthy people. The dog indicated which stool sample was from the cancer patient. This was repeated a total of 38 times. Out of the 38 tests, the dog correctly identified the cancer sample 37 times. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies cancer more than one fifth of the time. The p-value is less than .001. Assuming it was a well-designed study, use a significance level of .05 to make a decision.

a. Reject the null hypothesis and conclude that the dog correctly identifies cancer more than one fifth of the time.
b. There is enough statistical evidence to prove that the dog correctly identifies cancer more than one fifth of the time.
c. Do not reject the null hypothesis and conclude there is no evidence that the dog correctly identifies cancer more than one fifth of the time.

7F

27. Bob did a study where 40 subjects were randomly assigned to two groups (20 per group). He performed a study, analyzed the data, and found a *p*-value when testing if the mean difference between the groups was statistically significant. Imagine that Bob did a new study where 200 students were randomly assigned to two groups (100 per group). Assume that the observed mean difference for the new study is the same as the observed mean difference in original study. How would the *p*-value for new study (100 per group) compare to the *p*-value for original study (20 per group)?

a. It would be the same as the original *p*-value.
b. It would be smaller than the original *p*-value.
c. It would be larger than the original *p*-value.

8F

28. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. 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 from the Department of Natural Resources took a random

sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?

a. A random sample of a sample size of 100 with a sample mean of 12.1.
b. A random sample of a sample size of 36 with a sample mean of 11.5.
c. A random sample of a sample size of 100 with a sample mean of 11.5.
d. A random sample of a sample size of 36 with a sample mean of 12.1.

9F

29. A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes ($p = 0.287$). What does this mean?

a. The distribution of the GPAs for male and female scholarship athletes are identical except for 28.7% of the athletes.
b. The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287 or larger, assuming that there is a difference between the means.
c. There is a 28.7% chance that a pair of randomly chosen male and female scholarship athletes would have a significant difference assuming that there is no difference.
d. There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between males and females as that observed in the sample assuming that there is no difference.

10F

30. Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached.
The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean?

a. The increase in test scores could occur just by chance alone even if coaching really has no effect.
b. The increase in test scores makes no difference in getting into college since it is not statistically significant.
c. The study was badly designed because they did not have equal numbers of coached and not-coached students.

11F

31. A researcher is interested if there is a significant difference in the average number of hours watching television between male and female 8th grade students in the US. After gathering and analyzing the data, the researcher found that the difference in the average number of hours between the two groups was 4.37 hours/week. The researcher conducted a test to verify if this difference in means was statistically significant and found a p-value of 0.001. What would be the correct conclusion the researcher needs to make?

a. The difference between groups in the average number of hours watching television did happen by chance because the p-value is so small.
b. The difference between groups in the average number of hours watching television is *not* statistically significant because the p-value is too small.
c. There is evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there is *no* difference.
d. There is *not* strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there is a difference.

12F

32. A research article reports the result of a new drug test. The drug is hypothesized to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article reports a *p*-value of 0.04 in the analysis section.

Which option below presents the correct interpretations of this p-value?

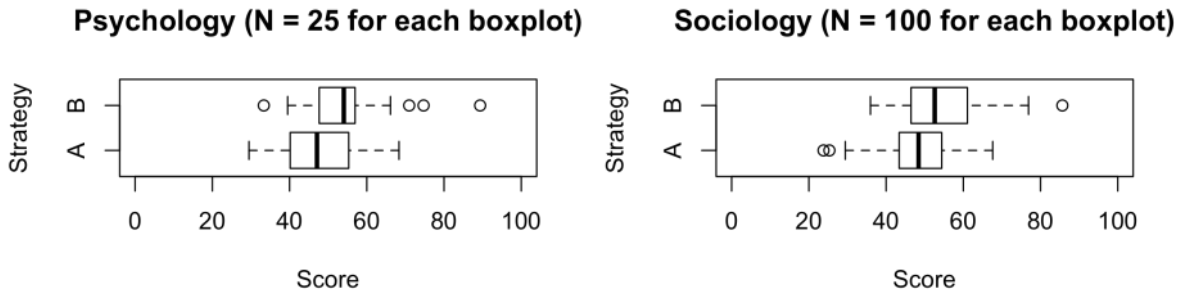
a. We conclude that the new drug is not effective because there is only a .04 probability that the drug is more effective than the current treatment.
b. We conclude that the new drug is effective because a result like they found, or more extreme, would only happen 4% of the time if the drug was not effective.
c. We conclude that the new drug is effective because there is only a 4% chance that it's not.
d. We conclude that the new drug is not effective because the difference in the proportion of macular degeneration patients with vision loss between the two treatments is only 0.04.

13F

33. Two experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100). The two different experiments were conducted with students who were enrolled in two different subject areas: psychology and sociology.

Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence

supporting the claim “one strategy is better than the other”? Why?



a. Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment.
b. Psychology, because there are more outliers in strategy B from the Psychology experiment, indicating that strategy B did not work well in that course.
c. Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment.
d. Sociology, because the sample size is larger in the Sociology experiment. This will produce a more accurate estimate of the difference between strategies.

14F

34. An engineer designs a new light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of 40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. A significant result for such a small difference would occur because:

a. The new design had more variability than the previous design.
b. The sample size for the new design is very large.
c. The mean of 1,200 for the previous design is large.

1E

35. The Pew Research Center surveyed 2,076 American adult cell phone users chosen at random in 2013. The sample percent of adult cell phone users who access the internet on their phones was 60%. The 95% confidence interval was 58% to 62%. What is this interval attempting to estimate?

a. The average number of American adult cell phone users who access the internet on their phones in 2013.
b. The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
c. The percent of all American adult cell phone users who access the internet on their phones in 2013.
d. For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.

2E

36. In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

a. We know that 37% of veterans in the *sample* have been divorced at least once.
b. We know that 37% of veterans in the *population* have been divorced at least once.
c. We can say with 95% confidence that 37% of veterans in the *sample* have been divorced at least once.
d. We can say with 95% confidence that 37% of veterans in the *population* have been divorced at least once.

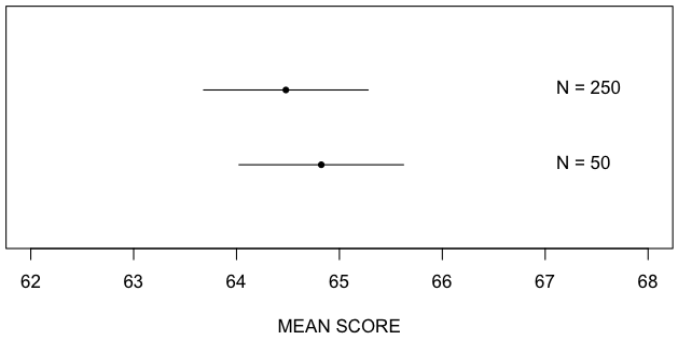
4E

37. Think about factors that might affect the width of a confidence interval. A confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval.

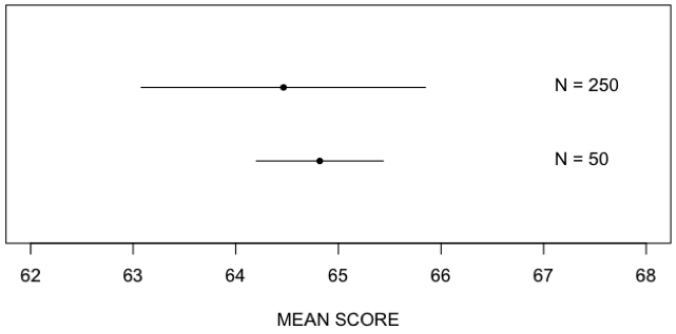
Imagine that two different random samples of test scores are drawn from a population of thousands of test scores. The first sample includes 250 test scores and the second sample includes 50 test scores. A 95% confidence interval for the population mean is constructed using each of the two samples.

Which set of confidence intervals (below) represents the two confidence intervals that would be constructed?

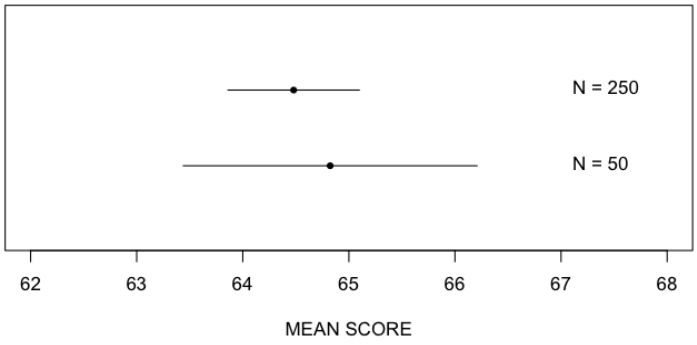
A.



B.



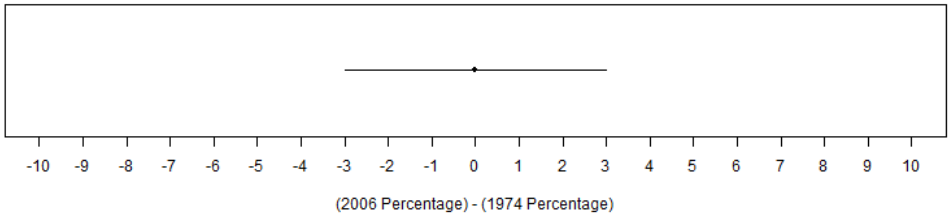
C.



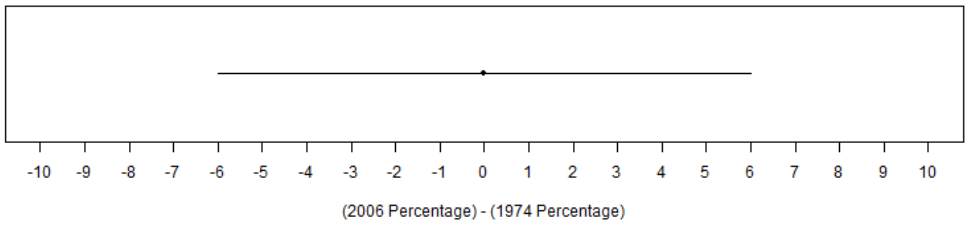
3E

38. A citizens' survey reported that in 1974, 64.5% of adults in a given state in the U.S. favored capital punishment, while in 2006 this percentage was 68.5%. To see if support for capital punishment has increased, a p-value for a test for difference in proportions is .011. Which of the following graphs represents a plausible 95% confidence interval for the difference in proportion of people who favored capital punishment between 1974 and 2006?

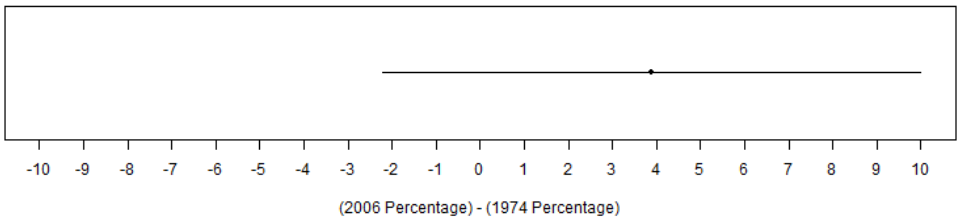
A



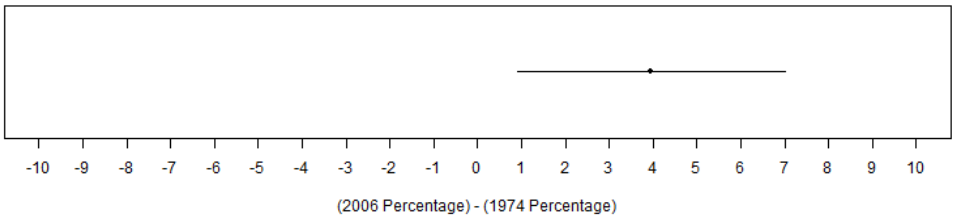
B



C



D



1G

39. Two students are flipping fair coins and recording whether or not the coin landed heads up. One student flips a coin 25 times and the other student flips a coin 125 times. Which student is more likely to get 48% to 52% of their coin flips heads up?

a. The student who flips the coin 25 times because the less flips that are made will increase the likelihood of a result of 50% heads up.
b. The student who flips the coin 125 times because the more flips that are made will increase the likelihood of a result of 50% heads up.
c. Neither student is more likely because the flipping of the coin is random and therefore you cannot predict the outcome of the flips.

2G

40. According to the National Cancer Institute, the probability of a man in the United States developing prostate cancer at some point during his lifetime is 0.15. What does the statistic, 0.15, mean in the context of this report from the National Cancer Institute?

a. For all men living in the United States, approximately 15% will develop prostate cancer at some point in their lives.
b. If you randomly selected a male in the United States there is a 15% chance that he will develop prostate cancer at some point in his life.
c. In a random sample of 100 men in the United States, 15 men will develop prostate cancer.
d. Both *a* and *b* are correct.

3G

41. The Gopher 5 is a cash lotto game in Minnesota. To play, players pick five numbers from 1 to 47. Each number can only be used once. The numbers are listed in numerical order, *not necessarily the order in which they were selected*. A player wins the Gopher 5 Jackpot if all five numbers chosen by that player match the five winning numbers chosen randomly by a computer. Here are four sets of five numbers that players have chosen for the Gopher 5:

Set 1: 5 – 10 – 15 – 20 – 25
Set 2: 1 – 13 – 25 – 31 – 42
Set 3: 10 – 16 – 24 – 25 – 40
Set 4: 1 – 2 – 3 – 4 – 5

Which of these sets of numbers is less likely to win the Gopher 5?

a. Set 1
b. Set 2
c. Set 3
d. Set 4
e. None of the above.

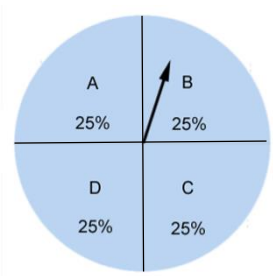
4G

42. A game company created a little plastic dog that can be tossed in the air. It can land either with all four feet on the ground, lying on its back, lying on its right side, or lying on its left side. However, the company does not know the probability of each of these outcomes. Which of the following methods is most appropriate to estimate the probability of each outcome?

a. Because there are four possible outcomes, assign a probability of 1/4 to each outcome.
b. Toss the plastic dog many times and see what percent of the time each outcome occurs.
c. Simulate the data using a model that has four equally likely outcomes.

5G

43. Consider a spinner shown below that has the letters from *A* to *D*.



Joan used the spinner 10 times and each time she wrote down the letter that the spinner landed on. When she looked at the results, she saw that the letter *B* showed up 5 times out of the 10 spins. Now she doubts the fairness of the spinner because it seems like she got too many *B*s.
A statistician wants to set up a probability model to examine how often the result of 5 B's out of 10 spins could happen with a fair spinner just by chance alone. Which of the following is the best probability model for the statistician to use?

a. The probability for each letter is the same - 1/4 for each letter.
b. The probability for letter B is 1/2 and the other three letters each have probability of 1/6.
c. The probability for letter B is 1/2 and the probabilities for the other letters sum to 1/2.

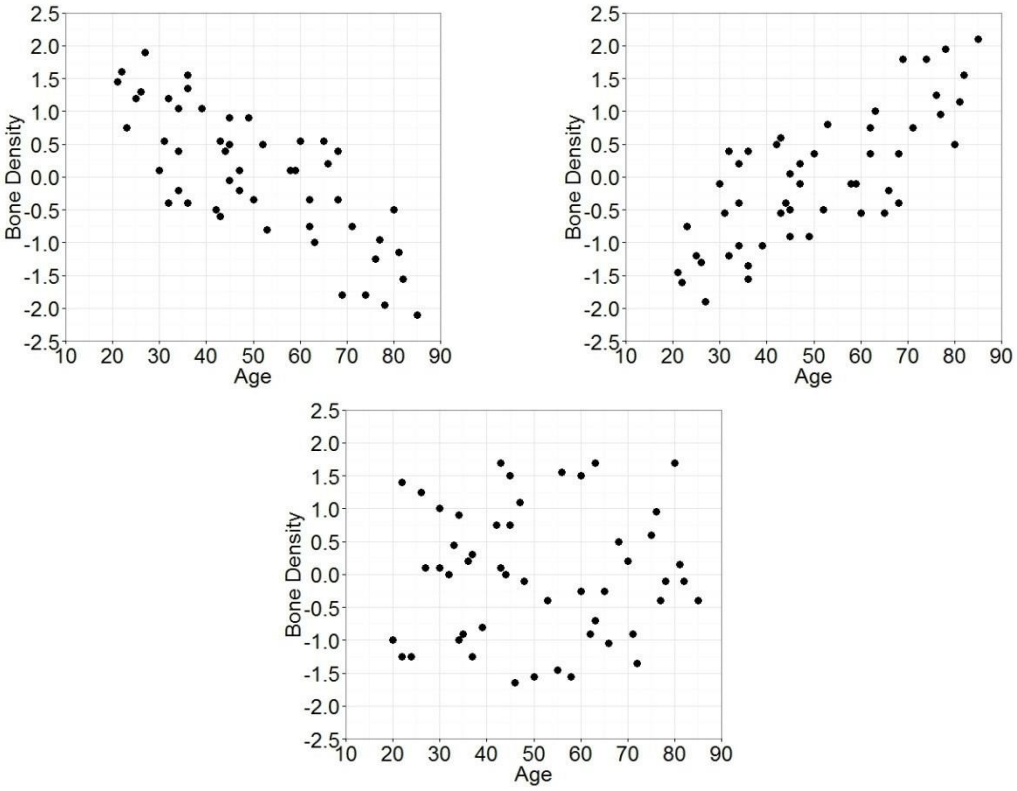
6G

44. Five faces of a fair die are painted black, and one face is painted white. The die is rolled six times.
Which of the following results is more likely?

a. Black side up on five of the rolls; white side up on the other roll
b. Black side up on all six rolls
c. *a* and *b* are equally likely

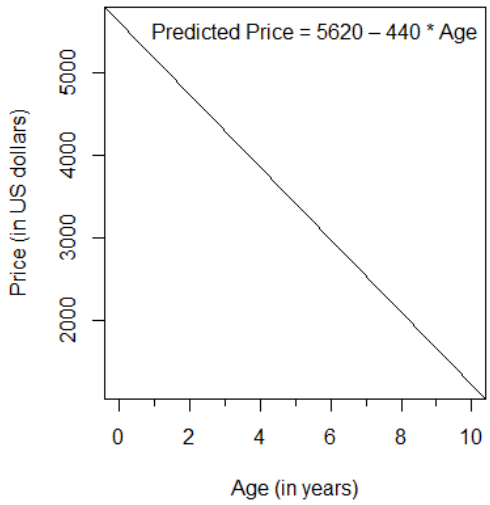
1H

45. Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?



2H

46. A statistics student gathered data on a large numbers of cars of a particular model, from new cars to those that were up to 10 years old. Using the data on car ages (in years) and car prices (in US dollars) he found a linear relationship and produced the following regression equation and plot of the regression equation:



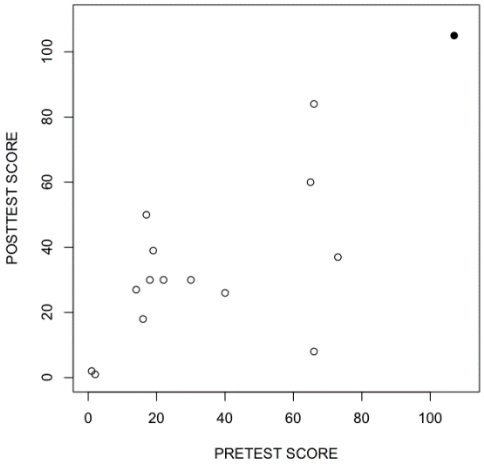
A friend asked him to use regression to predict the price of a 5 year-old model of this car. Which of the following methods can be used to provide an estimate?

a. Locate the point on the line that corresponds to an age of 5 and read off the corresponding value on the y axis.
b. Substitute an age of 5 in the equation and solve for "Predicted Price".
c. Both of these methods are correct.
d. Neither of these methods is correct.

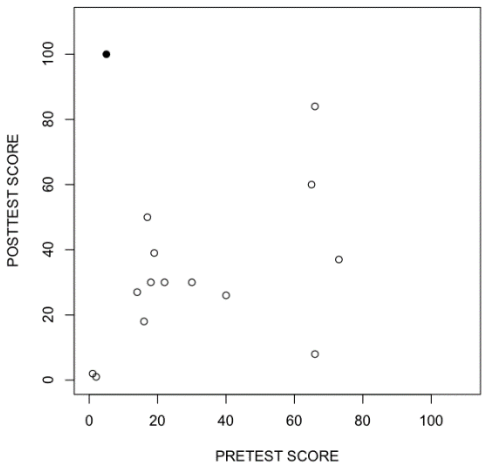
3H

47. On the first day of her statistics class, Dr. Andrew gave students a pretest to determine their statistical knowledge. At the end of the course, he gave students the exact same test. Dr. Andrew constructed the scatterplot (below on the left) between students' pretest and posttest scores. The solid point in the upper right corner of the scatterplot represents the pretest and posttest scores for Nicola. It turns out that Nicola's pretest score was actually 5 (not 100 as previously recorded), and her posttest score was 100. Nicola's scores were corrected and a new scatterplot was constructed (below on the right).

Scatterplot with Nicola's Incorrect Scores



Scatterplot with Nicola's Correct Scores

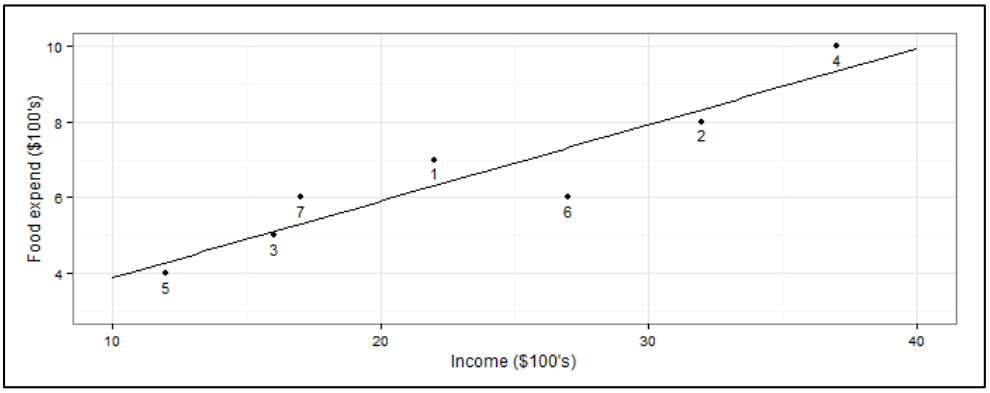


How would you expect the strength of the correlation between the pretest and posttest scores for the new scatterplot with Nicola's actual scores (above, right) to compare to the strength of the relationship for the original scatterplot (above, left)?

a. The new correlation would be weaker than the original correlation.
b. The new correlation would be stronger than the original correlation.
c. The new correlation would have the same strength as the original correlation.

5H

48. A random sample of 7 households was obtained, and information on their income and food expenditures for this year was collected. Below is a scatterplot of these data with the regression line superimposed:



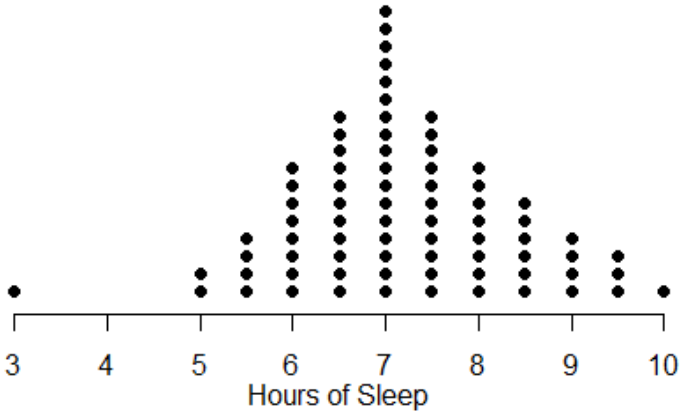
The regression equation for these data is: *predicted expend = 1.869 + 0.202 (income)*
After further analysis, the researcher found out that Point 4 was data from another year and decided to exclude it. What would be the regression equation after excluding Point 4?

a. predicted expend = 1.869 + 0.234 (income)
b. predicted expend = 1.869 + 0.202 (income)
c. predicted expend = 2.550 + 0.164 (income)
d. predicted expend = 1.821 + 0.234 (income)

C3 - Final version of the REALI assessment

REPRESENTATION OF DATA

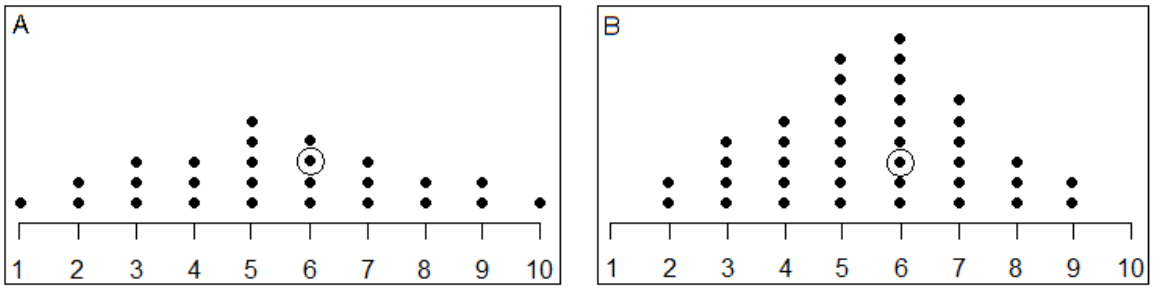
1. A high school teacher is concerned with whether her students are getting enough sleep. She asked each one of her students to report the number of hours slept the previous night. A plot of the results is given below.



She would like to report her findings to the school principal using appropriate statistical language to describe and interpret the distribution in context. Select the most appropriate statement that she could use to summarize the results.

a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five.
b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
c. Many students seem to be getting 7 hours of sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.
d. The distribution of hours of sleep is somewhat normal, with an outlier at 3. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.

2. Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights for 39 random samples of 3 pebbles each, with all mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in Figure A and the dot circled in Figure B? Please select the best answer from the list below.



a. No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
b. Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6.
c. Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B is the mean weight of 3 pebbles.

3. A teacher keeps track of the time it took her students to complete a particular exam (in minutes). These times are recorded in the table below.

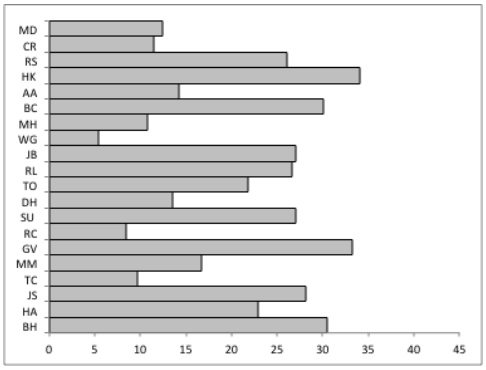
Student	Time
BH	30.5
HA	22.9
JS	28.1
TC	9.7
MM	16.7
GV	33.3
RC	8.5

Student	Time
SU	27.0
DH	13.6
TO	21.8
RL	26.7
JB	27.0
WG	5.4
MH	10.8

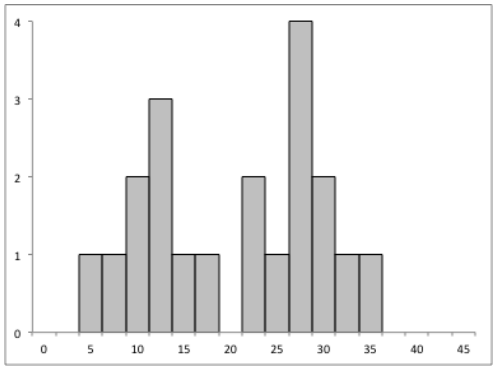
Student	Time
BC	30.1
AA	14.3
HK	34.1
RS	26.1
CR	11.5
MD	12.5

Each of the graphs below presents a valid representation of the time taken to complete the exam. Which of the graphs is the most appropriate display of the distribution of the times, in that the graph allows the teacher to describe the shape, center, and variability of the completion times?

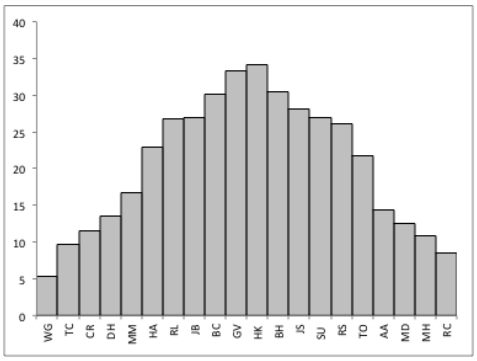
A.



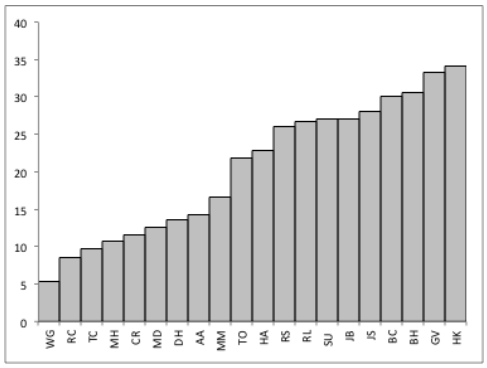
C.



B.



D.

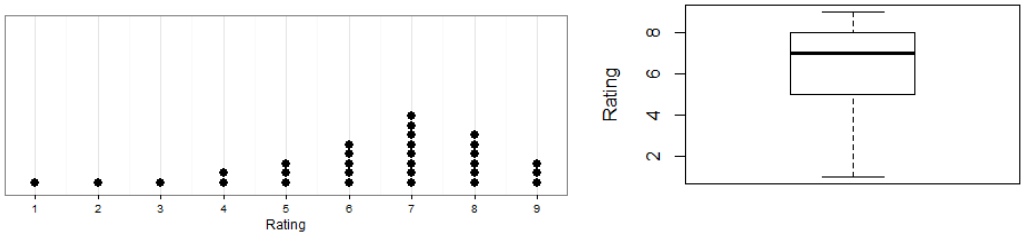


4. An instructor gave his introductory statistics students a survey. One of the questions read, "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. After analyzing the answers from the students, the instructor interpreted the data as follows:

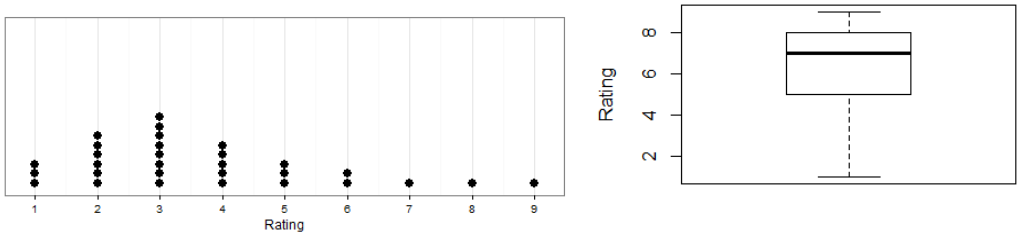
> "A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding."

The instructor asked two of his students to create a graphical representation of the data, based on his interpretation above. Beth created a dotplot and Allan created a boxplot. Which dotplot/boxplot pair best aligns with the description given by the instructor?

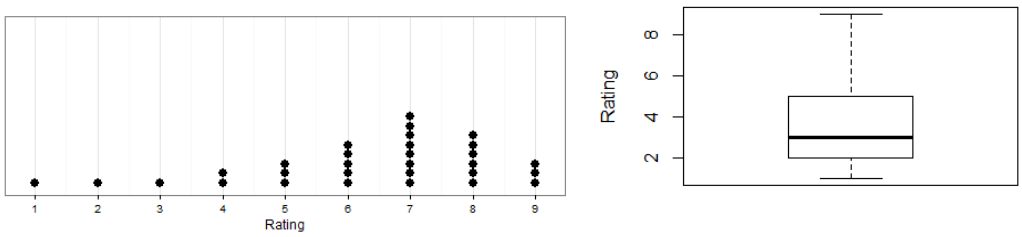
A



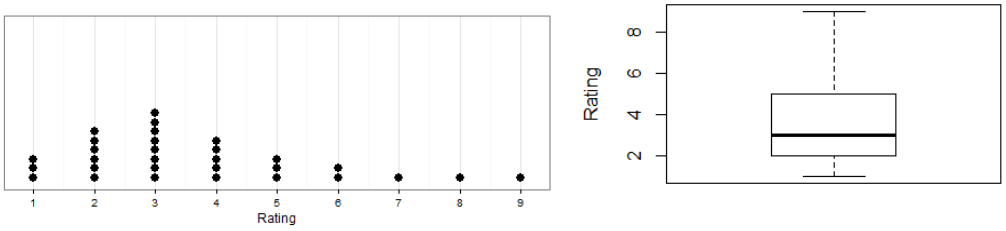
B



C

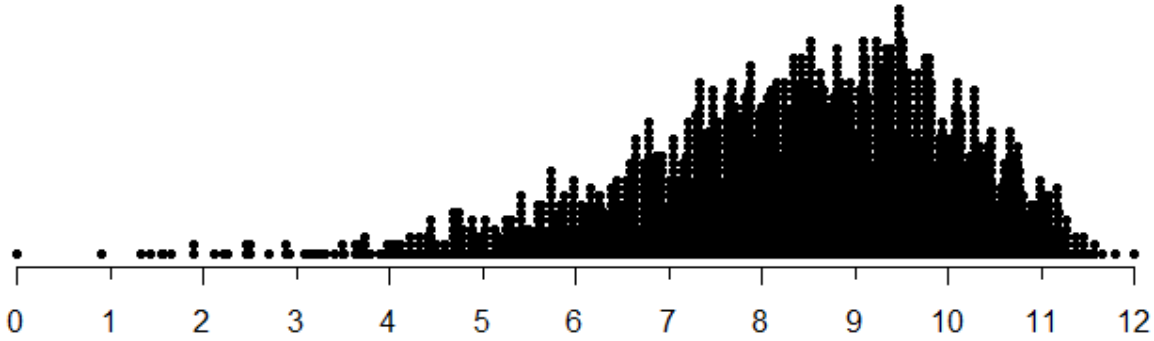


D



MEASURES OF CENTER

5. Which of the following intervals is *most* likely to include the mean of the distribution below?



a. 4 to 6
b. 7 to 9
c. 10 to 12

6. According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the average?

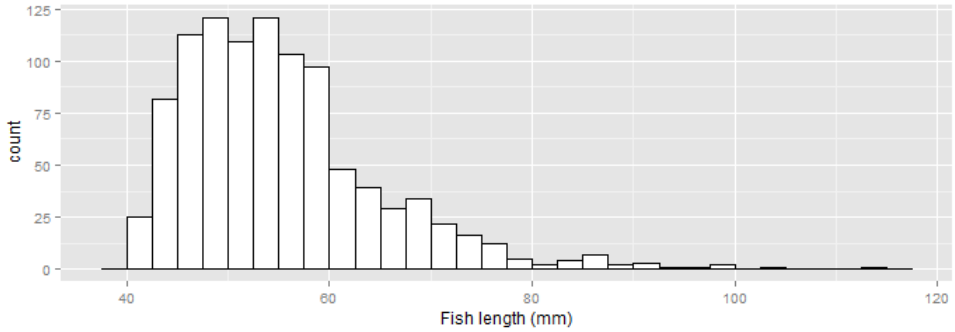
a. For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700.
b. For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700.
c. For all dog owners in this sample, about half were above $1,700 and about half were below $1,700.
d. For most owners, the first-year costs for owning a large-sized dog is $1,700.

7. In 2011, it was reported that the mean home price in the Hamptons (New York) increased by 20% within a single year, while the median home price decreased by 2% during that same year. Which of the following is the best explanation for this occurrence?

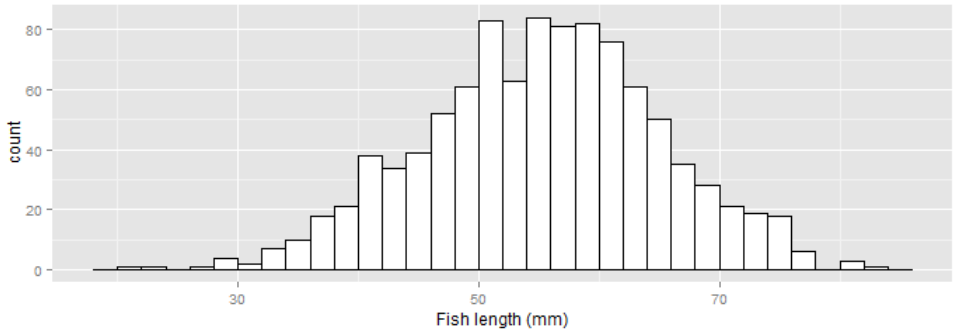
a. The price of most homes in the Hamptons decreased and more homes were sold in the Hamptons that year.
b. The reporters made an error in presenting the results; if the mean home price increases, the median home price must also increase.
c. Most of the homes in the Hamptons decreased in price and a small number of homes had large increases in price.

8. A study examined the length of a certain species of fish from one lake. A random sample of 1000 fish had mean length of 52.2mm and median length of 57.4mm. Which of the following histograms is most likely to represent the data from this study?

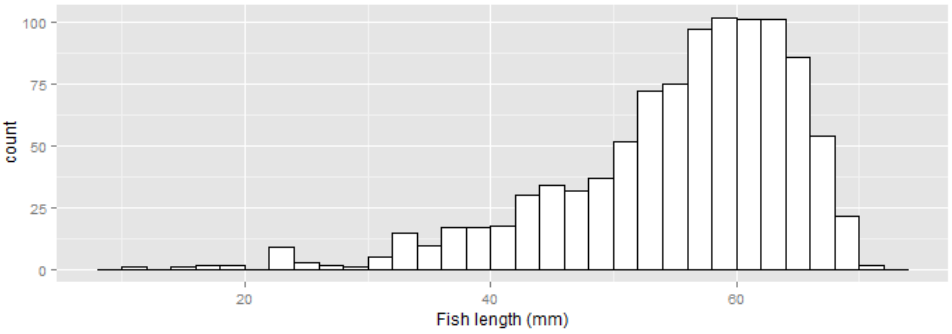
A



B



C



MEASURES OF VARIATION

9. Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quiz scores was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?

a. All of the individual scores are one point apart.
b. The difference between the highest and lowest score is 1 point.
c. The difference between the upper and lower quartile is 1 point.
d. A typical distance of a score from the mean is 1 point.

10. A teacher gives a science test with 15 questions. For each question, 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 to be -2.30. What do we know?

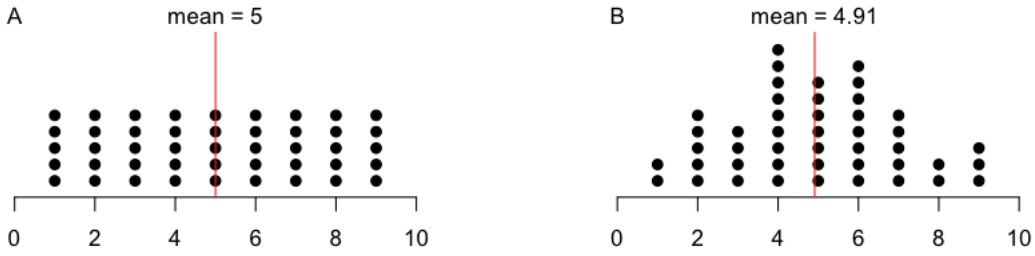
a. The standard deviation was calculated incorrectly.
b. Most students received negative scores.
c. Most students scored below the mean.
d. None of the above.

11. Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following commute times, in minutes, for 5 days of travel on each route.

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

a. The City Route
b. The Country Route
c. Neither route is better than the other.

12. Indicate which distribution has the larger standard deviation.



a. *A* has a larger standard deviation than *B*.
b. *B* has a larger standard deviation than *A*.
c. Both distributions have the same standard deviation.

13. A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level?

a. Observational study
b. Randomized experiment
c. Survey study

14. The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results. Identify also the sample from that population.

a. The population is all American adults in 2013. The sample is the 21% of American adults that have had an email or social networking account compromised.
b. The population is the 1,002 American adults surveyed. The sample is all American adults in 2013.
c. The population is all American adults in 2013. The sample is the 1,002 American adults surveyed.

15. CNN conducted a quick vote poll with a random sample of 5,581 Americans on September 19, 1999 to determine "What proportion of Americans still think that the Miss America pageant is relevant today?" For the sample, 1,192 people voted yes and 4,389 people voted no. Identify the statistic and parameter of interest.

a. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the 5,581 Americans who took part in the survey.
b. The statistic is the 5,581 Americans who took part in the survey and the parameter is all Americans.
c. The statistic is the proportion of all Americans who think the pageant is still relevant and the parameter is the sample proportion of people who voted yes (1192/5581= .214).
d. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the proportion of all Americans who think the pageant is still relevant.

16. A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.
Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

a. Invalid, because the sample is only 1/50th of the size of the population.
b. Invalid, because not all viewers had the opportunity to respond to the poll.
c. Invalid, because the sample may not be representative of the population.

17. Researchers conducted a survey of 1,000 randomly selected adults in the United States and found a strong, positive, statistically significant correlation between income and the number of containers the adults reported recycling in a typical week.
Can the researchers conclude that higher income causes more recycling among U.S. adults? Select the best answer from the following options.

a. No, the sample size is too small to allow causation to be inferred.
b. No, the lack of random assignment does not allow causation to be inferred.
c. Yes, the statistically significant result allows causation to be inferred.
d. Yes, the sample was randomly selected, so causation can be inferred.

18. A sportswriter wants to know the extent to which football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?

a. This is a simple random sample. It will give an accurate estimate.
b. Because the sample is so small, it will *not* give an accurate estimate.
c. Because all fans had a chance to be asked, it will give an accurate estimate.
d. The sampling method is biased. It will *not* give an accurate estimate.

19. A researcher wants to answer the following research question: Do sleep apnea patients look more attractive after receiving a new treatment (continuous positive airway pressure therapy)? Twenty patients with sleep apnea were given the new treatment. Pictures of the patients were taken before the treatment as well as two months after the treatment. An independent rater was shown the before and after pictures for each patient in a random order and asked to determine which picture the patient looked more attractive in. Seventy percent (70%) of the patients were rated as more attractive in the after picture. Is there a need to conduct a hypothesis test to determine whether the treatment is effective, or could the researcher just use the sample statistic (70%) as evidence of the effectiveness?

a. The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%; therefore, the result did not happen by chance.
b. The researcher should conduct a hypothesis test because a hypothesis test is always appropriate.
c. The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.

20. One hundred student-athletes attended a summer camp to train for a particular track race. All 100 student-athletes followed the same training program in preparation for an end-of-camp race. Fifty of the student-athletes were randomly assigned to additionally participate in a weight-training program along with their normal training (the training group). The other 50 student-athletes did not participate in the additional weight-training program (the non-training group). At the end of the summer camp, all 100 student-athletes ran the same race and their individual times (in seconds) were recorded.

The mean speed of the training group was 44 seconds, and the mean speed of the non-training group was 66 seconds. The standard deviation for the non-training group was 20 seconds. Consider the following possible values for the standard deviation of the training group. Which of these values would produce the strongest evidence of a difference in means between the two groups?

a. 10 seconds
b. 20 seconds
c. 30 seconds

21. The Amherst H. Wilder Foundation conducts surveys with homeless people in Minnesota. In their 2012 survey of 4,465 homeless adults, one question asked was "In the last 30 nights, how many nights have you spent outside, in a vehicle or vacant building, or some other place not intended for housing?" One research question that the surveyors asked was "Is there a difference between males and females with regards to the average number of nights spent in a place not intended for housing?" Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question?

a. There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.

b. There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
c. There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
d. There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.

22. Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up different scents. An experiment was conducted with a dog in Japan who was trained to smell bowel cancer in stool samples. In a test, the dog was presented with five stool samples; one from a cancer patient and four from healthy people. The dog indicated which stool sample was from the cancer patient. This was repeated a total of 38 times. Out of the 38 tests, the dog correctly identified the cancer sample 37 times. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies cancer more than one fifth of the time. The p-value is less than .001. Assuming it was a well- designed study, use a significance level of .05 to make a decision.

a. Reject the null hypothesis and conclude that the dog correctly identifies cancer more than one fifth of the time.
b. There is enough statistical evidence to prove that the dog correctly identifies cancer more than one fifth of the time.
c. Do not reject the null hypothesis and conclude there is no evidence that the dog correctly identifies cancer more than one fifth of the time.

23. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. 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 from the Department of Natural Resources took a random sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?

a. A random sample of a sample size of 100 with a sample mean of 12.1.
b. A random sample of a sample size of 36 with a sample mean of 11.5.
c. A random sample of a sample size of 100 with a sample mean of 11.5.
d. A random sample of a sample size of 36 with a sample mean of 12.1.

24. Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached.
The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean?

a. The increase in test scores could occur just by chance alone even if coaching really has no effect.
b. The increase in test scores makes no difference in getting into college since it is not statistically significant.
c. The study was badly designed because they did not have equal numbers of coached and not-coached students.

25. A researcher is interested if there is a significant difference in the average number of hours watching television between male and female 8th grade students in the US. After gathering and analyzing the data, the researcher found that the difference in the average number of hours between the two groups was 4.37 hours/week. The researcher conducted a test to verify if this difference in means was statistically significant and found a p-value of 0.001. What would be the correct conclusion the researcher needs to make?

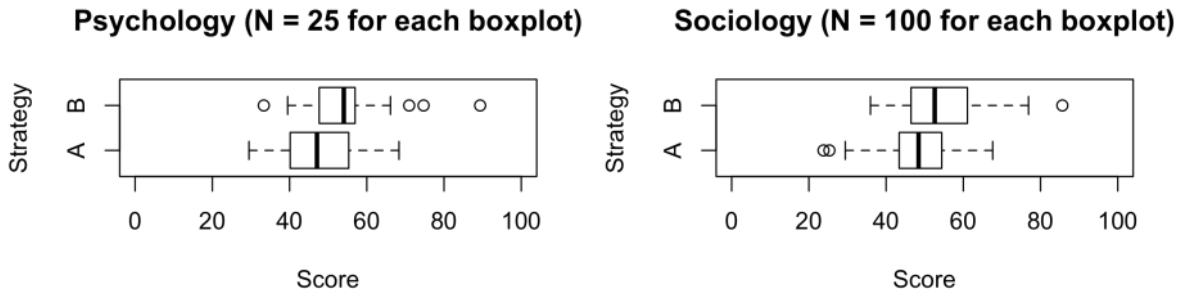
a. The difference between groups in the average number of hours watching television did happen by chance because the p-value is so small.
b. The difference between groups in the average number of hours watching television is *not* statistically significant because the p-value is too small.
c. There is evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there is *no* difference.
d. There is *not* strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there is a difference.

26. A research article reports the result of a new drug test. The drug is hypothesized to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article reports a *p*-value of 0.04 in the analysis section.
Which option below presents the correct interpretations of this p-value?

a. We conclude that the new drug is not effective because there is only a .04 probability that the drug is more effective than the current treatment.
b. We conclude that the new drug is effective because a result like they found, or more extreme, would only happen 4% of the time if the drug was not effective.
c. We conclude that the new drug is effective because there is only a 4% chance that it's not.
d. We conclude that the new drug is not effective because the difference in the proportion of macular degeneration patients with vision loss between the two treatments is only 0.04.

27. Two experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100). The two different experiments were conducted with students who were enrolled in two different subject areas: psychology and sociology.

Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence supporting the claim "one strategy is better than the other"? Why?



a. Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment.
b. Psychology, because there are more outliers in strategy B from the Psychology experiment, indicating that strategy B did not work well in that course.
c. Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment.
d. Sociology, because the sample size is larger in the Sociology experiment. This will produce a more precise estimate of the difference between strategies.

28. An engineer designs a new light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of 40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. A significant result for such a small difference would occur because:

a. The new design had more variability than the previous design.
b. The sample size for the new design is very large.
c. The mean of 1,200 for the previous design is large.

29. The Pew Research Center surveyed 2,076 American adult cell phone users chosen at random in 2013. The sample percent of adult cell phone users who access the internet on their phones was 60%. The 95% confidence interval was 58% to 62%. What is this interval attempting to estimate?

a. The average number of American adult cell phone users who access the internet on their phones in 2013.
b. The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
c. The percent of all American adult cell phone users who access the internet on their phones in 2013.
d. For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.

30. In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

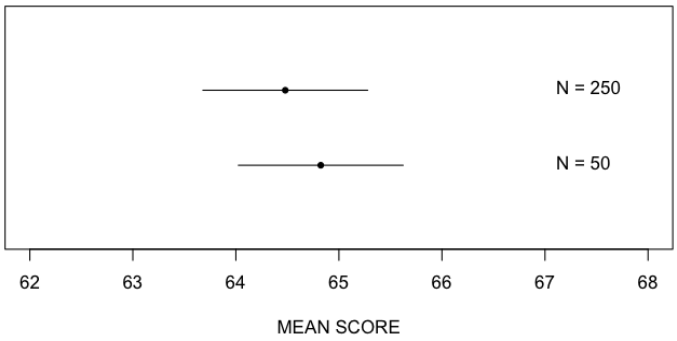
a. We can say that 37% of veterans in the *sample* have been divorced at least once.
b. We can say that 37% of veterans in the *population* have been divorced at least once.
c. We can say that 95% of veterans in the *sample* have been divorced at least once.
d. We can say that 95% of veterans in the *population* have been divorced at least once.

31. Think about factors that might affect the width of a confidence interval. A confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval.

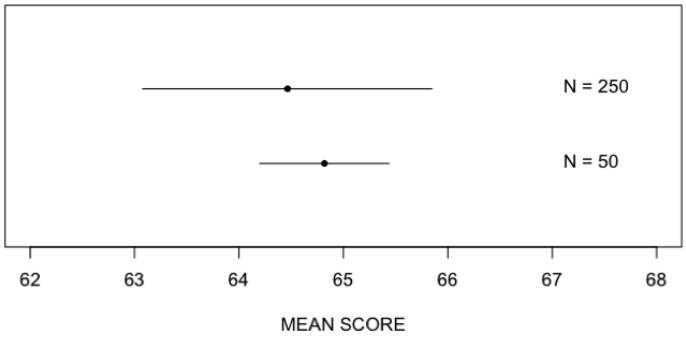
Imagine that two different random samples of test scores are drawn from a population of thousands of test scores. The first sample includes 250 test scores and the second sample includes 50 test scores. A 95% confidence interval for the population mean is constructed using each of the two samples.

Which set of confidence intervals (below) represents the two confidence intervals that would be constructed?

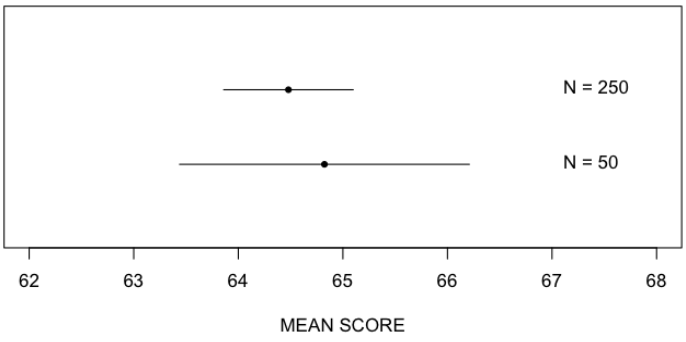
A.



B.

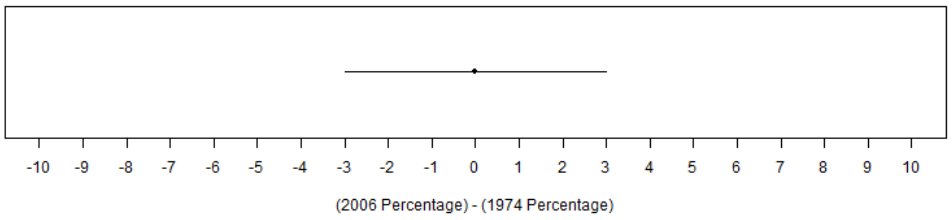


C.

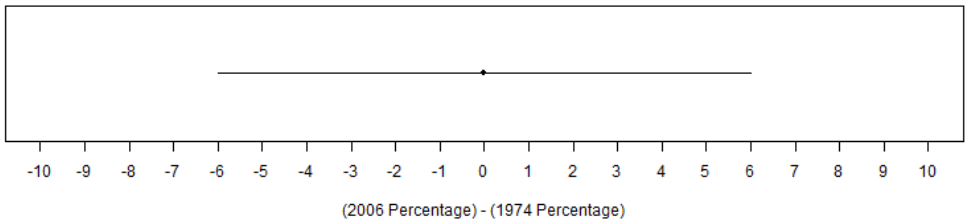


32. A citizens' survey reported that in 1974, 64.5% of adults in a given state in the U.S. favored capital punishment, while in 2006 this percentage was 68.5%. To see if support for capital punishment has increased, a p-value for a test for difference in proportions is .011. Which of the following graphs represents a plausible 95% confidence interval for the difference in proportion of people who favored capital punishment between 1974 and 2006?

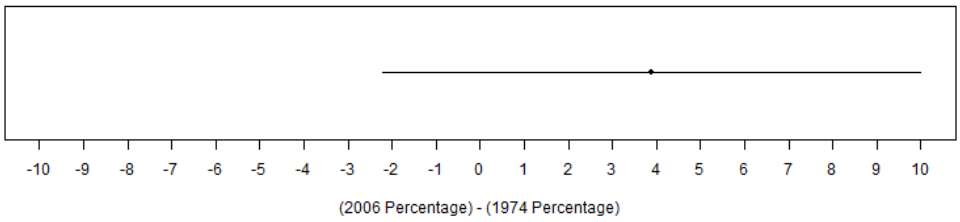
A



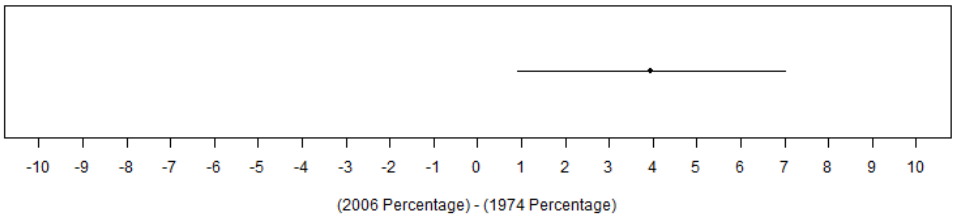
B



C



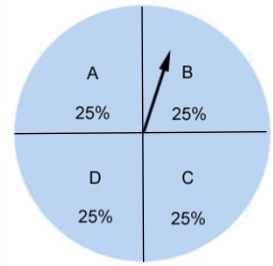
D



35. A game company created a little plastic dog that can be tossed in the air. It can land either with all four feet on the ground, lying on its back, lying on its right side, or lying on its left side. However, the company does not know the probability of each of these outcomes. Which of the following methods is most appropriate to estimate the probability of each outcome?

a. Because there are four possible outcomes, assign a probability of 1/4 to each outcome.
b. Toss the plastic dog many times and see what percent of the time each outcome occurs.
c. Simulate the data using a model that has four equally likely outcomes.

36. Consider a spinner shown below that has the letters from *A* to *D*.

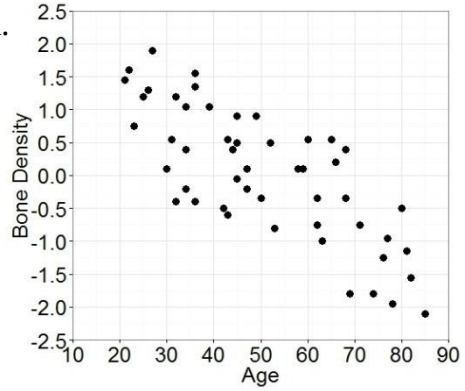


Joan used the spinner 10 times and each time she wrote down the letter that the spinner landed on. When she looked at the results, she saw that the letter *B* showed up 5 times out of the 10 spins. Now she doubts the fairness of the spinner because it seems like she got too many *B*s.
A statistician wants to set up a probability model to examine how often the result of 5 B's out of 10 spins could happen with a fair spinner just by chance alone. Which of the following is the best probability model for the statistician to use?

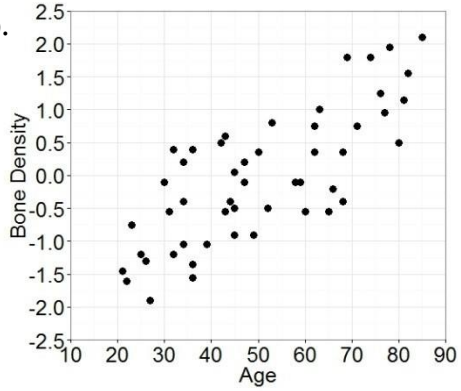
a. The probability for each letter is the same - 1/4 for each letter.
b. The probability for letter B is 1/2 and the other three letters each have probability of 1/6.
c. The probability for letter B is 1/2 and the probabilities for the other letters sum to 1/2.

37. Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?

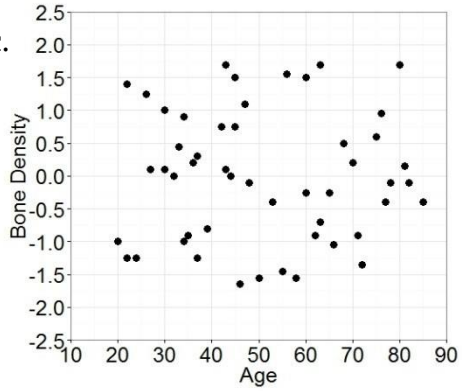
a.



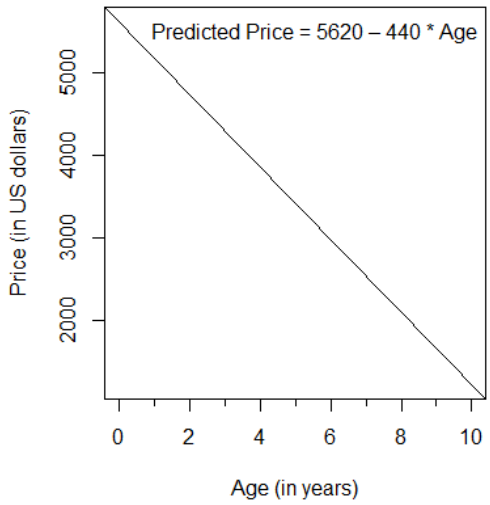
b.



c.



38. A statistics student gathered data on a large numbers of cars of a particular model, from new cars to those that were up to 10 years old. Using the data on car ages (in years) and car prices (in US dollars) he found a linear relationship and produced the following regression equation and plot of the regression equation:

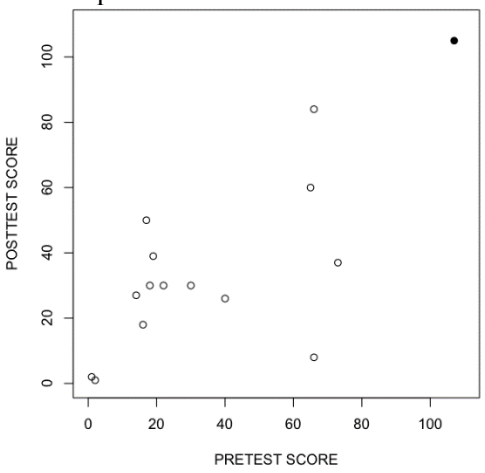


A friend asked him to use regression to predict the price of a 5 year-old model of this car. Which of the following methods can be used to provide an estimate?

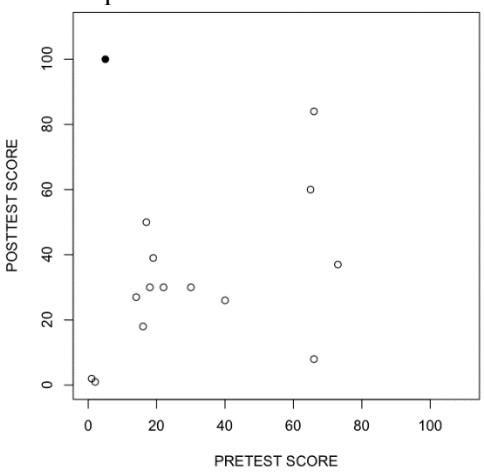
a. Locate the point on the line that corresponds to an age of 5 and read off the corresponding value on the y axis.
b. Substitute an age of 5 in the equation and solve for "Predicted Price".
c. Both of these methods are correct.
d. Neither of these methods is correct.

39. On the first day of her statistics class, Dr. Andrew gave students a pretest to determine their statistical knowledge. At the end of the course, he gave students the exact same test. Dr. Andrew constructed the scatterplot (below on the left) between students' pretest and posttest scores. The solid point in the upper right corner of the scatterplot represents the pretest and posttest scores for Nicola. It turns out that Nicola's pretest score was actually 5 (not 100 as previously recorded), and her posttest score was 100. Nicola's scores were corrected and a new scatterplot was constructed (below on the right).

Scatterplot with Nicola's Incorrect Scores



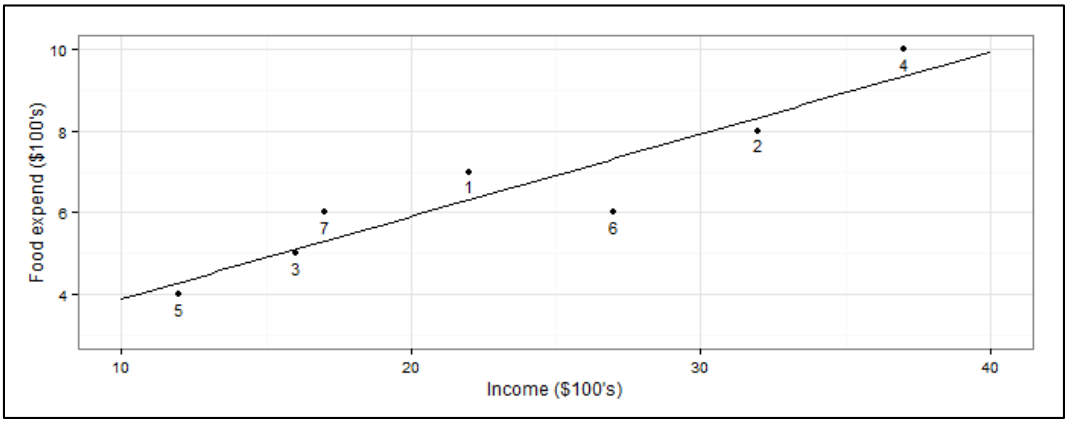
Scatterplot with Nicola's Correct Scores



How would you expect the strength of the correlation between the pretest and posttest scores for the new scatterplot with Nicola's actual scores (above, right) to compare to the strength of the relationship for the original scatterplot (above, left)?

a. The new correlation would be weaker than the original correlation.
b. The new correlation would be stronger than the original correlation.
c. The new correlation would have the same strength as the original correlation.

40. A random sample of 7 households was obtained, and information on their income and food expenditures for this year was collected. Below is a scatterplot of these data with the regression line superimposed:



The regression equation for these data has an intercept of 1.869 and a slope of 0.202:

Predicted expend = 1.869 + 0.202 (income).

After further analysis, the researcher found out that Point 4 was data from another year and decided to exclude it. What would be true about the regression equation after excluding Point 4?

a. Both the *intercept* and the *slope* will increase.
b. Both the *intercept* and the *slope* will decrease.
c. The *intercept* will decrease and the *slope* will increase.
d. The *intercept* will increase and the *slope* will decrease.
e. Both the *intercept* and the *slope* will remain the same.

Appendix D: Expert Review Correspondence

D1 - Email Invitation – phase 1

Dear Professor X,

I am currently doing research on important learning goals in introductory statistics courses and I am writing to request your assistance. My research is part of my PhD dissertation at the University of Minnesota, where I am a doctoral candidate in the Department of Educational Psychology with a concentration in Statistics Education. I am working with my advisers, Joan Garfield and Andrew Zieffler.

I am requesting your help with this project because of your expertise in the area of statistics education. The task you will be asked to perform is classifying 50 items into two groups.

If you are willing to participate, please inform if you are available at any of the times below so that we can meet and I can give you the material and observe you while you do the classification. The estimated time for the classification is around 45 minutes.

Monday (10/12/2015): any time before 11am or after 1pm.
Tuesday (10/13/2015): any time before 10am
Wednesday (10/14/2015): any time before 2pm or after 5pm.

Sincerely,
Anelise G. Sabbag

D2 - Email Invitation – phase 2

Dear Professor X,

I am doctoral student beginning my dissertation research project which focuses on important learning goals in introductory statistics courses at the tertiary levels. I am writing to request your assistance in this project which involves assessing and distinguishing these outcomes. I am in the Statistics Education graduate program at the University of Minnesota, where I am working with my advisers, Joan Garfield and Andrew Zieffler.

I am asking for your help because of your expertise in the area of statistics education and in the area of defining student learning outcomes such as statistical literacy and statistical reasoning. I am developing an assessment that measures these learning goals and I will gather data to empirically distinguish them and determine whether or not they are separate constructs or related to each other.

If you agree to participate in my research, the task you will be asked to perform is classifying 50 items into two groups. This will be done in a word document and will take no more than 45 minutes.

If you agree to participate as an expert reviewer, I will send you the 50 items and the instructions for classifying them on *October 24*th. The turnaround for the classification will *November 9th* (this will give a little more than 2 weeks). Please feel free to ask me any questions that you have. I sincerely hope that you will be able to contribute to my research.

Please let me know if you are able to participate.

Sincerely,
Anelise G. Sabbag
Doctoral candidate
Department of Educational Psychology
University of Minnesota

D3 - Expert Review Form

EXPERT REVIEW

The task you are being asked to perform is classifying 50 items into two groups. The definitions of each group are:

GROUP 1
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).

GROUP 2
Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

Instructions

- Read the definition of each group carefully (definitions are also presented on every page).
- Read each item and think about the *minimum* understanding needed by a student to answer the item correctly.
- Classify each item as belonging to one of the two groups. In your classification, please use *only* the definitions given above.

Both definitions refer to "statistical concepts". Example of "statistical concepts" are: mean, standard deviation, confidence intervals, graphical representations, etc…

There is a space for you to write any concerns or comments you might have about your classification or about the item itself (this part is optional.)

If you have any questions about the classification process please email me at **sabb0013@umn.edu** or call me (612-859-5955).

Thank you very much for your help,

Anelise G Sabbag
Doctoral Candidate in Statistics Education
Department of Educational Psychology
University of Minnesota
sabb0013@umn.edu

REALI ASSESSMENT

4G

A game company created a little plastic dog that can be tossed in the air. It can land either with all four feet on the ground, lying on its back, lying on its right side, or lying on its left side. However, the company does not know the probability of each of these outcomes. Which of the following methods is most appropriate to estimate the probability of each outcome?

a. Since there are four possible outcomes, assign a probability of 1/4 to each outcome.
b. Toss the plastic dog many times and see what percent of the time each outcome occurs.
c. Simulate the data using a model that has four equally likely outcomes.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

6G

Five faces of a fair die are painted black, and one face is painted white. The die is rolled six times. Which of the following results is more likely?

a. Black side up on five of the rolls; white side up on the other roll
b. Black side up on all six rolls
c. *a* and *b* are equally likely

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

2G

According to the National Cancer Institute, the probability of a man in the United States developing prostate cancer at some point during his lifetime is 0.15. What does the statistic, 0.15, mean in the context of this report from the National Cancer Institute?

a. For all men living in the United States, approximately 15% will develop prostate cancer at some point in their lives.
b. If you randomly selected a male in the United States there is a 15% chance that he will develop prostate cancer at some point in his life.
c. In a random sample of 100 men in the United States, 15 men will develop prostate cancer.
d. Both *a* and *b* are correct.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

3G

The Gopher 5 is a cash lotto game in Minnesota. To play, players pick five numbers from 1 to 47. Each number can only be used once. The numbers are listed in numerical order (not necessarily the order in which they were selected). A player wins the Gopher 5 Jackpot if all five numbers chosen by that player match the five winning numbers chosen randomly by a computer. Here are four sets of five numbers that players have chosen for the Gopher 5:

Set 1: 5 – 10 – 15 – 20 – 25
Set 2: 1 – 13 – 25 – 31 – 42
Set 3: 10 – 16 – 24 – 25 – 40
Set 4: 1 – 2 – 3 – 4 – 5

Which of these sets of numbers is less likely to win the Gopher 5?

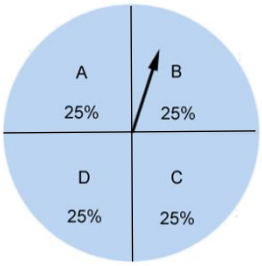
a. Set 1
b. Set 2
c. Set 3
d. Set 4
e. None of the above.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

5G

Consider a spinner shown below that has the letters from *A* to *D*.



Joan used the spinner 10 times and each time she wrote down the letter that the spinner landed on. When she looked at the results, she saw that the letter *B* showed up 5 times out of the 10 spins. Now she doubts the fairness of the spinner because it seems like she got too many *B*s.
A statistician wants to set up a probability model to examine how often the result of 5 B's out of 10 spins could happen with a fair spinner just by chance alone. Which of the following is the best probability model for the statistician to use?

a. The probability for each letter is the same - 1/4 for each letter.
b. The probability for letter B is 1/2 and the other three letters each have probability of 1/6.
c. The probability for letter B is 1/2 and the probabilities for the other letters sum to 1/2.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

1G

Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 50 times and the other student flips a coin 100 times. Which student is more likely to get 48% to 52% of their coin flips heads up?

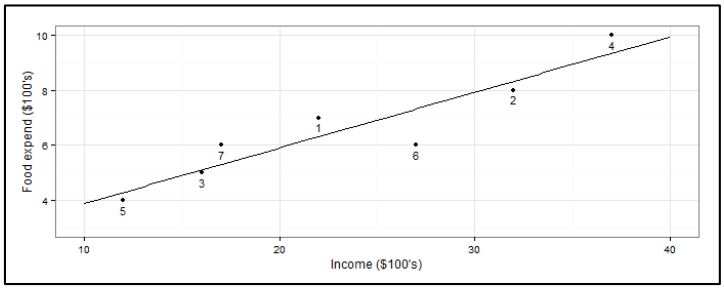
a. The student who flips the coin 50 times because the percent that are heads up is less likely to be exactly 50%.
b. The student who flips the coin 100 times because that student has more chances to get a coin flip that is heads up.
c. The student who flips the coin 100 times because the more flips that are made will increase the chance of approaching a result of 50% heads up.
d. Neither student is more likely because the flipping of the coin is random and therefore you cannot predict the outcome of the flips.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

5H

A random sample of 7 households was obtained, and information on their income and food expenditures for this year was collected. Below is a scatterplot of these data with the regression line superimposed:



The regression equation for these data is: ***predicted expend = 1.869 + 0.202 (income)***
After further analysis, the researcher found out that Point 4 was data from another year and decided to exclude it. What would be the regression equation after excluding Point 4?

a. predicted expend = 1.869 + 0.234 (income)
b. predicted expend = 1.869 + 0.202 (income)
c. predicted expend = 2.550 + 0.164 (income)
d. predicted expend = 1.821 + 0.234 (income)

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

6H

Past data was collected from students who take a certain statistics class, where test scores ranged from 0-100. The regression line relating the final exam score and the midterm exam score is:

Predicted final exam = 50 + 0.05(midterm)

What is a correct interpretation of the ***slope***?

a. A student who scored 0 on the final exam would be predicted to score 50 on the midterm exam.
b. As the score on the final increases by 1 point, it is predicted that the score on the midterm exam will increase by 0.05 point.
c. A student who scored 0 on the midterm would be predicted to score 50 on the final exam.
d. As the score on the midterm increases by 1 point, it is predicted that the score on the final exam will increase by 0.05 point.

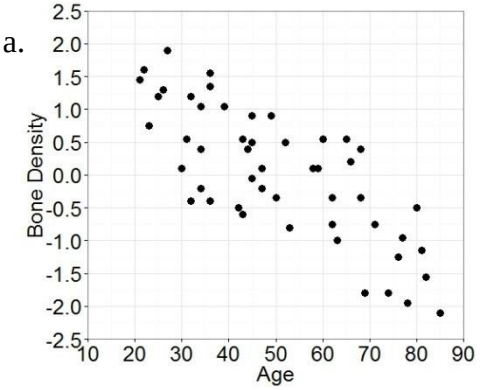
CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

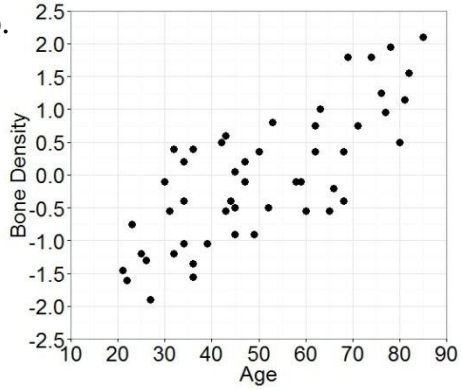
1H

Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?

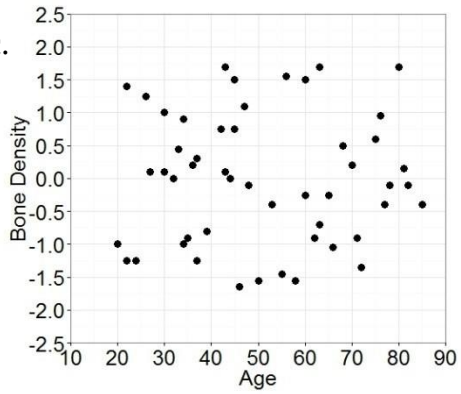
a.


Bone Density
Age
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
10
20
30
40
50
60
70
80
90

b.


Bone Density
Age
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
10
20
30
40
50
60
70
80
90

c.


Bone Density
Age
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
10
20
30
40
50
60
70
80
90

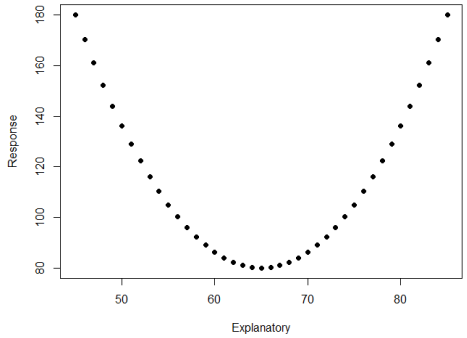
CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

1H

COMMENTS (OPTIONAL):

4H

The correlation between two variables is zero and the scatterplot for these variables is presented below.



Based on the correlation value and the plot above, how would you interpret the relationship between the explanatory variable and the response variable?

a. There is no relationship between the explanatory variable and the response variable.
b. Correlation = 0 indicates a perfect relationship between an explanatory variable and a response variable.
c. Correlation is not appropriate to quantify the strength of the relationship between these variables.

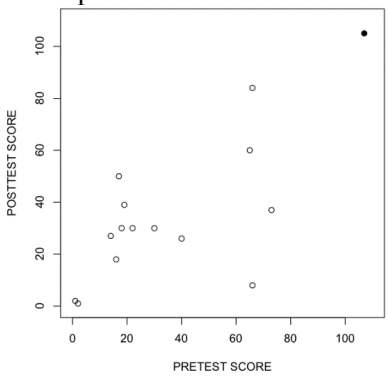
CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

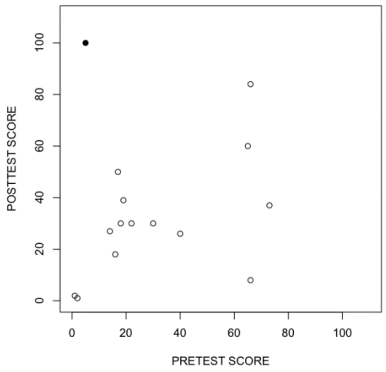
3H

On the first day of her statistics class, Dr. Smith gave students a pretest to determine their statistical knowledge. At the end of the course, she gave students the exact same test. Dr. Smith constructed the scatterplot (below on the left) between students' pretest and posttest scores. The solid point in the upper right corner of the scatterplot represents the pretest and posttest scores for John. It turns out that John's pretest score was actually 5 (not 100 as previously recorded), and his posttest score was 100. John's scores were corrected and a new scatterplot was constructed (below on the right).

Scatterplot with John's Incorrect Scores



Scatterplot with John's Correct Scores



How would you expect the strength of the correlation between the pretest and posttest scores for the new scatterplot with John's actual scores (above, right) to compare to the strength of the relationship for the original scatterplot (above, left)?

a. The new correlation would be weaker than the original correlation.
b. The new correlation would be stronger than the original correlation.
c. The new correlation would have the same strength as the original correlation.

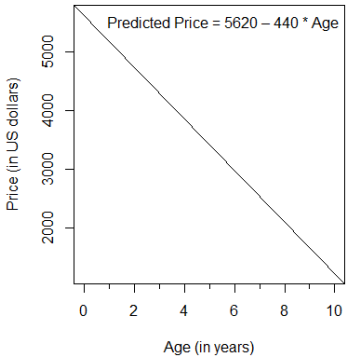
CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

3H

COMMENTS (OPTIONAL):

2H

A statistics student gathered data on a large numbers of cars of a particular model, from new cars to those that were up to 10 years old. Using the data on car ages (in years) and car prices (in US dollars) he found a linear relationship and produced the following regression equation and plot of the regression equation:



A friend asked him to use regression to predict the price of a 5 year-old model of this car. Which of the following methods can be used to provide an estimate?

a. Locate the point on the line that corresponds to an age of 5 and read off the corresponding value on the y axis.
b. Substitute an age of 5 in the equation and solve for "Predicted Price".
c. Both of these methods are correct.
d. Neither of these methods is correct.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

2H

COMMENTS (OPTIONAL):

3A

A teacher keeps track of the time it took her students to complete a particular exam (in minutes). These times are recorded in the table below.

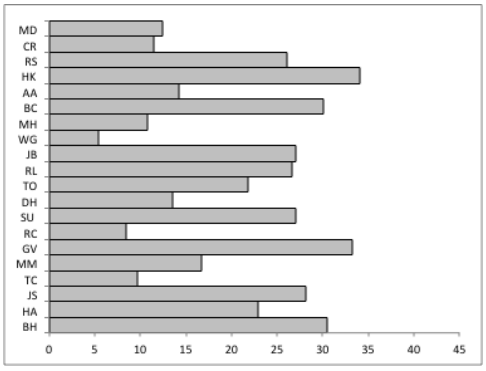
Student	Time
BH	30.5
HA	22.9
JS	28.1
TC	9.7
MM	16.7
GV	33.3
RC	8.5

Student	Time
SU	27.0
DH	13.6
TO	21.8
RL	26.7
JB	27.0
WG	5.4
MH	10.8

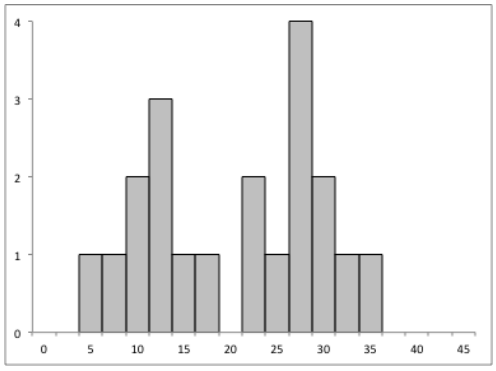
Student	Time
BC	30.1
AA	14.3
HK	34.1
RS	26.1
CR	11.5
MD	12.5

Each of the graphs below presents a valid representation of the time taken to complete the exam. Which of the graphs is the most appropriate display of the distribution of the times, in that the graph allows the teacher to describe the shape, center, and variability of the completion times?

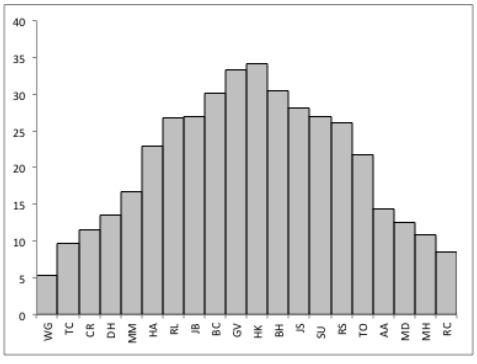
A.



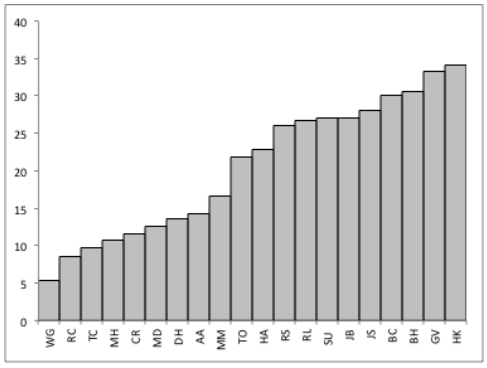
B.



C.



D.



3A

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

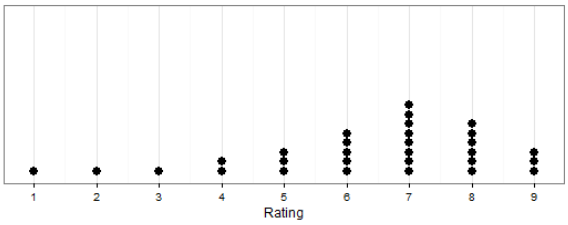
5A

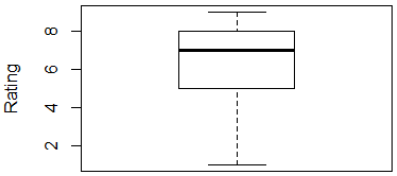
One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. After analyzing the answers from the students, the instructor interpret the data saying

> "**A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding."**

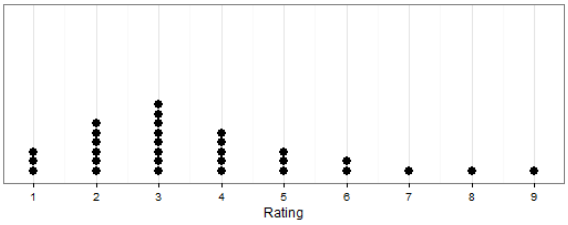
The instructor asked two of his students to create a graphical representation of the data, based on his interpretation above. Beth created a dotplot and Allan created a boxplot. Which dotplot/boxplot pair better aligns with the description given by the instructor?

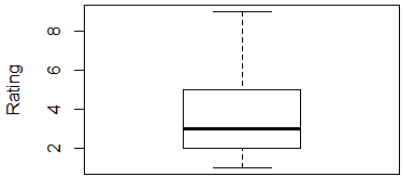
A



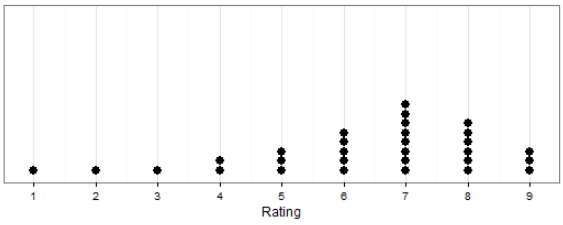


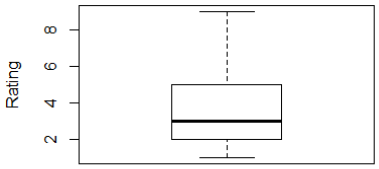
B



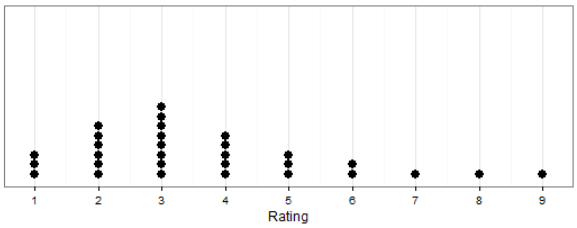


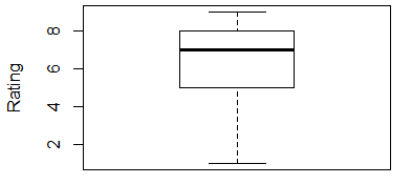
C





D



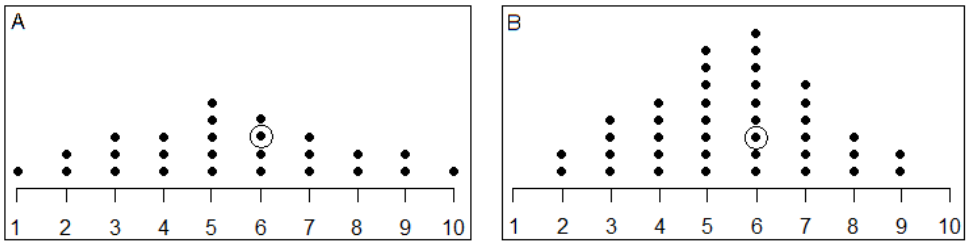


CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

2A

Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights for 39 random samples of 3 pebbles each, with all mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.



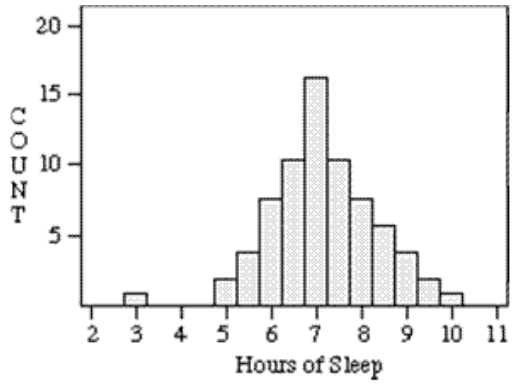
a. No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
b. Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6.
c. Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B represents the average weight of 3 pebbles.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

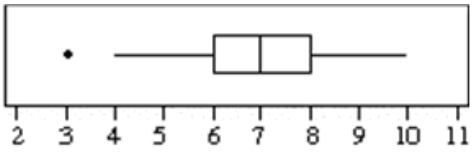
4A

The following graph shows a distribution of hours slept last night by a group of college students.

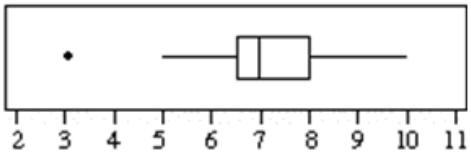


Which box plot seems to be graphing the same data as the histogram above?

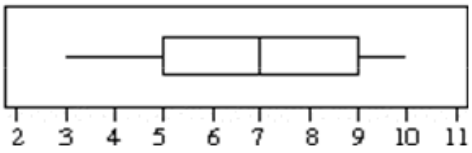
A



B



C



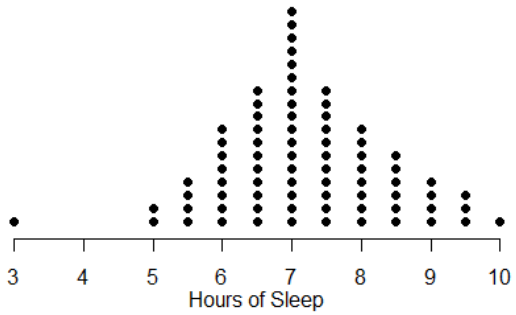
4A

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

1A

The following graph shows the distribution of hours slept the previous night by a group of college students.



Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.

a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five.
b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
c. Many students seem to be getting 7 hours of sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.
d. The distribution of hours of sleep is somewhat normal, with an outlier at 3. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

1A

COMMENTS (OPTIONAL):

7D

A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?

a. This is a simple random sample. It will give an accurate estimate.
b. Because the sample is so small, it will not give an accurate estimate.
c. Because all fans had a chance to be asked, it will give an accurate estimate.
d. The sampling method is biased. It will not give an accurate estimate.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

1D

The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population.

a. The population is all American adults in 2013. The sample is the 21% of American adults that have had an email or social networking account compromised.
b. The population is the 1,002 American adults surveyed. The sample is all American adults in 2013.
c. The population is all American adults in 2013. The sample is the 1,002 American adults surveyed.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

8D

A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that 'prescription medications only' was the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?

a. Yes, because medication patients lived longer.
b. No, because doctors chose the treatments.
c. Yes, because the study was a comparative experiment.
d. No, because the patients volunteered to be studied.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

5D

A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum

that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.
Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

a. Valid, because the sample size is large enough to represent the population.
b. Valid, because 67% is far enough above 50% to predict a majority vote.
c. Invalid, because the sample is too small given the size of the population.
d. Invalid, because the sample may not be representative of the population.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

6D

Researchers conducted a survey of 1,000 randomly selected adults in the United States and found a strong, positive, statistically significant correlation between income and the number of containers the adults reported recycling in a typical week.

Can the researchers conclude that higher income causes more recycling among U.S. adults? Select the best answer from the following options.

a. No, the sample size is too small to allow causation to be inferred.
b. No, the lack of random assignment does not allow causation to be inferred.
c. Yes, the statistically significant result allows causation to be inferred.
d. Yes, the sample was randomly selected, so causation can be inferred.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

3D

A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level?

a. Observational study
b. Randomized experiment

c. Survey

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

4D

A research study randomly assigned participants into two groups. One group was given Vitamin E to take daily. The other group received only a placebo pill. The research study followed the participants for eight years. After the eight years, the proportion of each group that developed a particular type of cancer was compared.
What is the primary reason that the study used random assignment?

a. To ensure that the groups are similar in all respects except for the level of Vitamin E
b. To ensure that a person doesn't know whether or not they are getting the placebo.

c. To ensure that the study participants are representative of the larger population.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

2D

CNN conducted a quick vote poll with a random sample of 5,581 Americans on September 19, 1999 to determine "What proportion of Americans still think that the Miss America pageant is still relevant today?" For the sample, 1,192 people voted yes and 4,389 people voted no. Identify the statistic and parameter of interest.

a. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the 5,581 Americans who took part in the survey.
b. The statistic is the 5,581 Americans who took part in the survey and the parameter is all Americans.
c. The statistic is the proportion of all Americans who think the pageant is still relevant and

the parameter is the sample proportion of people who voted yes (1192/5581= .214).

d. The statistic is the sample proportion of people who voted yes (1192/5581= .214) and the parameter is the proportion of all Americans who think the pageant is still relevant.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

3B

In 2011, it was reported that the mean home price in the Hamptons (New York) increased by 20% within a single year, while the median home price decreased by 2% during that same year. Which of the following is the best explanation for this occurrence?

a. The price of most homes in the Hamptons decreased and more homes were sold in the Hamptons that year.
b. The reporters made an error in presenting the results; if the mean home price increases, the median home price must also increase.
c. Most of the homes in the Hamptons decreased in price and a small number of homes had large increases in price.

CLASSIFICATION OF THE ITEM ABOVE

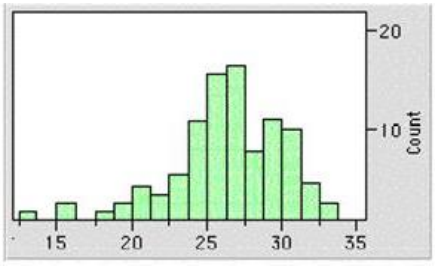
Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

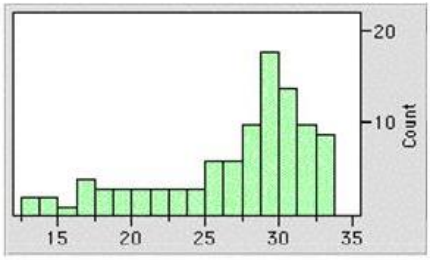
4B

A study examined the length of a certain species of fish from one lake. The plan was to take a random sample of 100 fish and examine the results. The mean length was 26.8mm, the median was 29.4mm, and the standard deviation was 5.0mm. Which of the following histograms is most likely to be the one for these data?

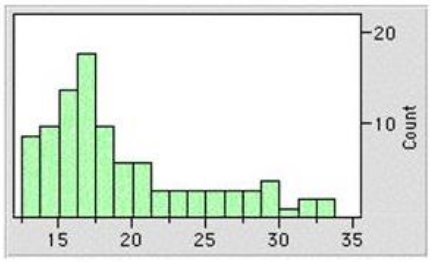
A



B



C



CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

4B

COMMENTS (OPTIONAL):

1B

According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the mean?

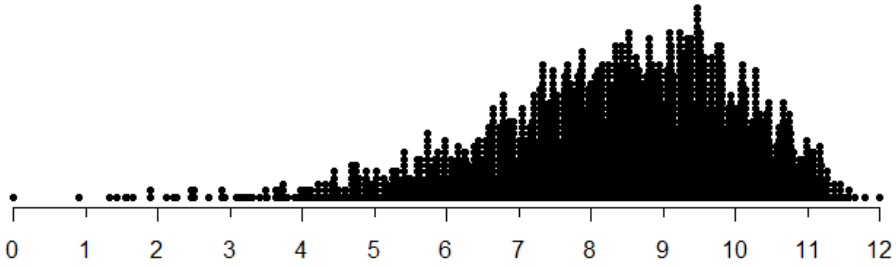
a. For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700.
b. For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700.
c. For all dog owners in this sample, about half were above $1,700 and about half were below $1,700.
d. For most owners, the first-year costs for owning a large-sized dog is $1,700.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

2B

Which of the following intervals is MOST likely to include the mean of the distribution below?



a. 6 to 7
b. 8 to 9
c. 9 to 10
d. 10 to 11

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

6B

The school committee of a small town wanted to determine the average number of children per household in their town. They divided the total number of children in the town by 50, the total

number of households. Which of the following statements must be true if the average children per household is 2.2?

a. Half the households in the town have more than 2 children.
b. More households in the town have 3 children than have 2 children.
c. There are a total of 110 children in the town.
d. There are 2.2 children in the town for every adult.
e. The most common number of children in a household is 2.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

12F

A research article reports the results of a new drug test. The drug is hypothesized to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article reports a *p*-value of 0.04 in the analysis section.
Which option below presents the correct interpretations of this p-value?

a. We conclude that the new drug is not effective because there is only a .04 probability that the drug is more effective than the current treatment.
b. We conclude that the new drug is effective because results like they found, or results even more favorable to the new drug, would only happen 4% of the time if the drug was not effective.
c. We conclude that the new drug is effective because there is only a 4% chance that it's not.
d. We conclude that the new drug is not effective because the difference in the proportion of macular degeneration patients with vision loss between the two treatments is only 0.04.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

8F

It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. 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 from the Department of Natural Resources took a random sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?

a. A random sample of a sample size of 100 with a sample mean of 12.1.
b. A random sample of a sample size of 36 with a sample mean of 11.5.
c. A random sample of a sample size of 100 with a sample mean of 11.5.
d. A random sample of a sample size of 36 with a sample mean of 12.1.

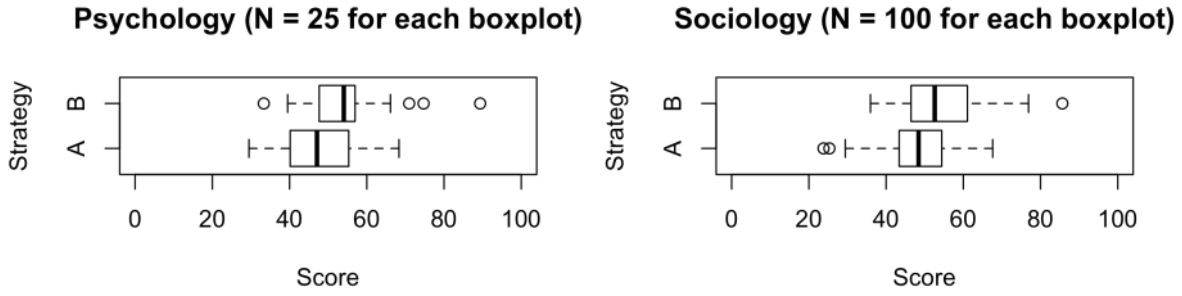
CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

13F

Two experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100). The two different experiments were conducted with students who were enrolled in two different subject areas: psychology and sociology.

Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence against the claim, "neither strategy is better than the other"? Why?



a. Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment.
b. Psychology, because there are more outliers in strategy B from the Psychology experiment, indicating that strategy B did not work well in that course.
c. Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment.
d. Sociology, because the sample size is larger in the Sociology experiment, which will produce a more accurate estimate of the difference between the two strategies.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

13F

COMMENTS (OPTIONAL):

7F

Bob did a study where 40 subjects were randomly assigned to two groups (20 per group). He performed a study, analyzed the data, and found a *p*-value when testing if the mean difference between the groups was statistically significant. Imagine that Bob did a new study where 200 students were randomly assigned to two groups (100 per group). Assume that the observed mean difference for the new study is the same as the observed mean difference in original study. How would the *p*-value for new study (100 per group) compare to the *p*-value for original study (20 per group)?

a. It would be the same as the original *p*-value.
b. It would be smaller than the original *p*-value.
c. It would be larger than the original *p*-value.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

5F

Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up

different scents. An experiment was conducted with a dog in Japan who was trained to smell bowel cancer in stool samples. In a test, the dog was presented with five stool samples; one from a cancer patient and four from healthy people. The dog indicated which stool sample was from the cancer patient. This was repeated a total of 38 times. Out of the 38 tests, the dog correctly identified the cancer sample 37 times. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies cancer more than one fifth of the time. The p-value is less than .001. Assuming it was a well- designed study, use a significance level of .05 to make a decision.

a. Reject the null hypothesis and conclude that the dog correctly identifies cancer more than one fifth of the time.
b. There is enough statistical evidence to prove that the dog correctly identifies cancer more than one fifth of the time.
c. Do not reject the null hypothesis and conclude there is no evidence that the dog correctly identifies cancer more than one fifth of the time.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

4F

A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain?

a. A large *p*-value.
b. A small *p*-value.
c. The magnitude of a *p*-value has no impact on statistical significance.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

11F

A researcher is interested if there is a significant difference in the average number of hours watching television between males and females 8th grade students in the US. After gathering and analyzing the data, the researcher found that the difference in the average number of hours between the two groups was 4.37 hours. The researcher conducted a test to verify if this difference in means was statistically significant and found a p-value of 0.001. What would be the correct conclusion the researcher needs to make?

a. The difference between groups in the average number of hours watching television did happen by chance because the p-value is so small.
b. The difference between groups in the average number of hours watching television is NOT statistically significant because the p-value is too small.
c. There IS strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there is NO **difference.**
d. There is NOT strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there IS a difference.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

3F

The Amherst H. Wilder Foundation conducts surveys with homeless people in Minnesota. In their 2012 survey of 4,465 homeless adults, one question asked was "In the last 30 nights, how many nights have you spent outside, in a vehicle or vacant building, or some other place not intended for housing?" One research question that the surveyors had was "Is there a difference between males and females with regards to the average number of nights spent in a place not intended for housing?" Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question?

a. There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
b. There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
c. There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
d. There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

14F

An engineer designs a new light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of

40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. A significant result for such a small difference would occur because:

a. The new design had more variability than the previous design.
b. The sample size for the new design is very large.
c. The mean of 1,200 for the previous design is large.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

2F

A researcher wants to answer the following research question: Do sleep apnea patients look more attractive after receiving a new treatment (continuous positive airway pressure therapy)? Twenty patients with sleep apnea were given the new treatment. Pictures of the patients were taken before the treatment as well as two months after the treatment. An independent rater was shown the before and after pictures for each patient in a random order and asked to determine which picture the patient looked more attractive in. Seventy percent (70%) of the patients were rated as more attractive in the after picture. Is there a need to conduct a hypothesis test to determine whether the treatment is effective, or could the researcher just use the sample statistic (70%) as evidence of the effectiveness?

a. The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%.
b. The researcher should conduct a hypothesis test because a hypothesis test is always appropriate.
c. The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

1F

The following situation models the logic of a hypothesis test. An electrician tests whether or

not an electrical circuit is good. The null hypothesis is that the circuit is good. The alternative hypothesis is that the circuit is not good. The electrician performs the test and decides to reject the null hypothesis. Which of the following statements is true?

a. The circuit is definitely not good and needs to be repaired.
b. The circuit is most likely not good, 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 might not be good.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

6F

One hundred student-athletes attended a summer camp to train for a particular track race. All 100 student-athletes followed the same training program in preparation for an end-of-camp race. Fifty of the student-athletes were randomly assigned to additionally participate in a weight-training program

along with their normal training (the training group). The other 50 student-athletes did not participate in the additional weight-training program (the non-training group). At the end of the summer camp, all 100 student-athletes ran the same race and their individual times (in seconds) were recorded.

The mean speed of the training group was 44 seconds, and the mean speed of the non-training group was 66 seconds. The standard deviation for the non-training group was 20 seconds. Consider the following possible values for the standard deviation of the training group. Which of these values would produce the strongest evidence of a difference between the two groups?

a. 10 seconds
b. 20 seconds
c. 30 seconds

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

9F

A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes ($P = 0.287$). What does this mean?

a. The distribution of the GPAs for male and female scholarship athletes are identical except for 28.7% of the athletes.
b. The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287.
c. There is a 28.7% chance that a pair of randomly chosen male and female scholarship athletes would have a significant difference assuming that there is no difference.
d. There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between male and female scholarship athletes as that observed in the sample assuming that **there is no** difference.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

10F

Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached.

The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean?

a. The sample sizes were too small to detect a true difference between the coached and not-coached students.
b. The observed difference between coached and not-coached students could occur just by chance alone even if coaching really has no effect.
c. The increase in test scores makes no difference in getting into college since it is not statistically significant.
d. The study was badly designed because they did not have equal numbers of coached and not-coached students.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

2E

In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

a. We know that 37% of veterans in the *sample* have been divorced at least once.
b. We know that 37% of veterans in the *population* have been divorced at least once.

c. We can say with 95% confidence that 37% of veterans in the *sample* have been divorced at least once.
d. We can say with 95% confidence that 37% of veterans in the *population* have been divorced at least once.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

1E

The Pew Research Center surveyed 2,076 American adult cell phone users chosen at random in 2013. The sample percent of adult cell phone users who access the internet on their phones was 60%. The 95% confidence interval was 58% to 62%. What is this interval attempting to estimate?

a. The average number of American adult cell phone users who access the internet on their phones in 2013.
b. The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
c. The percent of all American adult cell phone users who access the internet on their

phones in 2013.

d. For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.

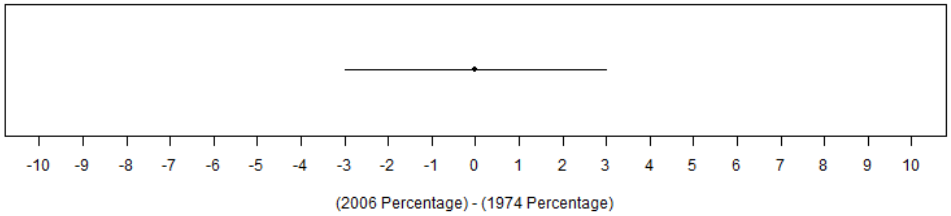
CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

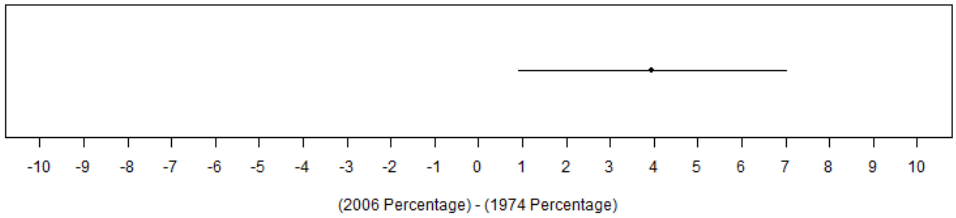
3E

A citizens' survey reported that in 1974, 64.5% of adults in a given state in the U.S. favored capital punishment, while in 2006 this percentage was 68.5%. To see if support for capital punishment has increased, a p-value for a test for difference in proportions is .011. Which of the following graphs represents a plausible 95% confidence interval for the difference in proportion of people who favored capital punishment between 1974 and 2006?

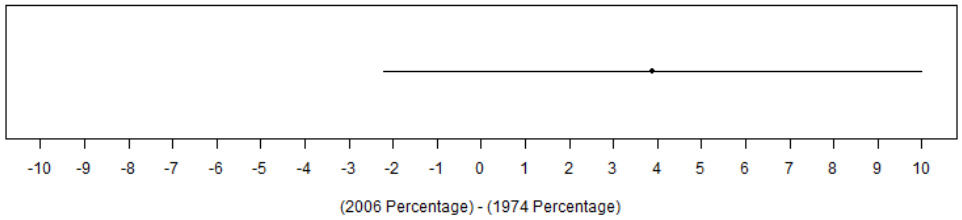
A


-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
(2006 Percentage) - (1974 Percentage)

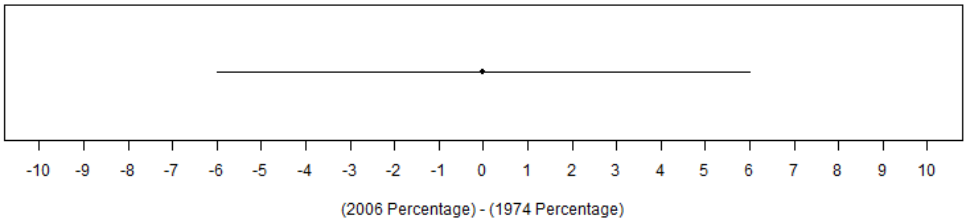
B


-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
(2006 Percentage) - (1974 Percentage)

C


-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
(2006 Percentage) - (1974 Percentage)

D


-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
(2006 Percentage) - (1974 Percentage)

3E

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

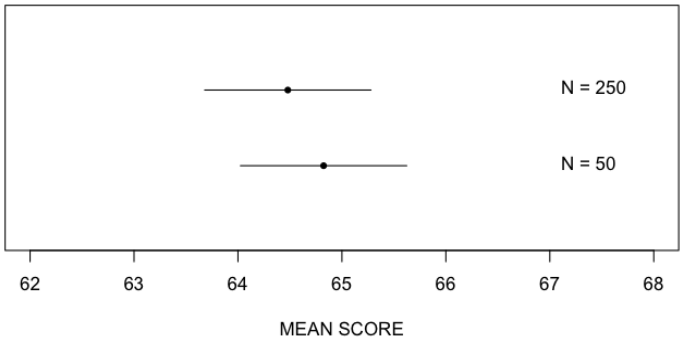
4E

Think about factors that might affect the width of a confidence interval. A confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval.

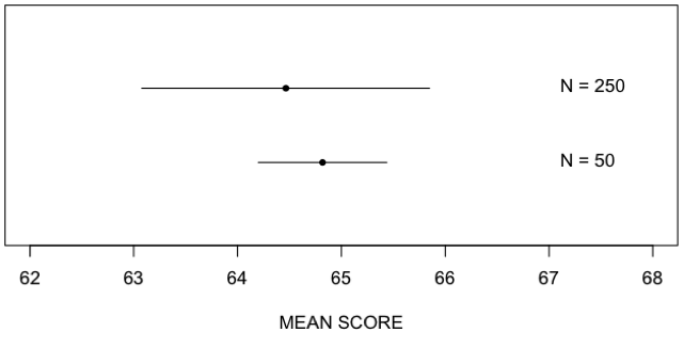
Imagine that two different random samples of test scores are drawn from a population of thousands of test scores. The first sample includes 250 test scores and the second sample includes 50 test scores. A 95% confidence interval for the population mean is constructed using each of the two samples.

Which set of confidence intervals (below) represents the two confidence intervals that would be constructed?

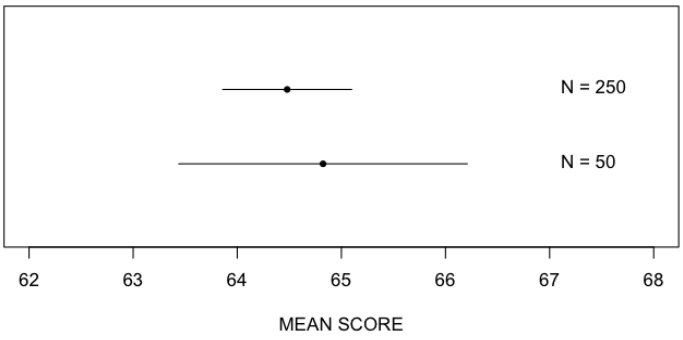
A.



B.



C.



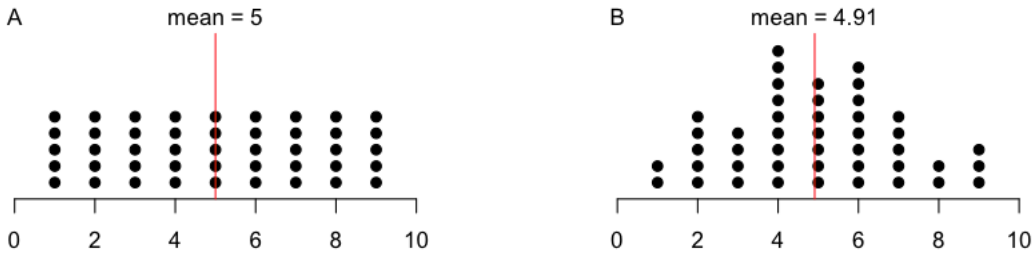
4E

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

4C

Indicate which distribution has the larger standard deviation.



a. A has a larger standard deviation than B.
b. B has a larger standard deviation than A.
c. Both distributions have the same standard deviation.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

2C

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 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 above.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

1C

Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quiz scores was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?

a. All of the individual scores are one point apart.
b. The difference between the highest and lowest score is 1 point.
c. The difference between the upper and lower quartile is 1 point.
d. A typical distance of a score from the mean is 1 point.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

3C

Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following data for 5 days of travel on each route.

Country Route - 17, 15, 17, 16, 18
City Route - 18, 13, 20, 10, 16

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

a. The Country Route, because the times are consistently between 15 and 18 minutes.
b. The City Route, because she can get there in 10 minutes on a good day and the average time is less than for the Country Route.
c. Because the times on the two routes have so much overlap, neither route is better than the other. She might as well flip a coin.

CLASSIFICATION OF THE ITEM ABOVE Please select which group the item belongs to by clicking one of the boxes below.	
GROUP 1 ☐	**GROUP 2** ☐
Items in GROUP 1 assess students' ability to *recall* a definition, *describe* or *interpret* basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will **not** require that students make connections between them (recall information will be sufficient).	Items in GROUP 2 assess students' ability to *make connections* among statistical concepts, *create mental representations* of statistical problems, or *explain relationships* between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, items from GROUP 2 require higher order thinking and higher cognitive load than items in GROUP 1.

COMMENTS (OPTIONAL):

D4 – Categorization of Items done by Experts

Item	Author Classification	Expert A	Expert B	Expert C	Expert D
Representation of data					
1A	S. Literacy	S. Reasoning	S. Literacy	S. Literacy	S. Literacy
2A	S. Literacy	S. Literacy	S. Literacy	S. Reasoning	S. Literacy
3A	S. Reasoning	S. Literacy	S. Literacy	S. Literacy	S. Literacy
4A	S. Reasoning	S. Reasoning	S. Literacy	S. Reasoning	S. Literacy
5A	S. Reasoning	S. Reasoning	S. Reasoning	S. Reasoning	S. Reasoning
Measures of center					
1B	S. Literacy	S. Literacy	S. Literacy	no time to answer	S. Literacy
2B	S. Literacy	S. Literacy	S. Literacy	no time to answer	S. Literacy
6B	S. Literacy	S. Literacy	S. Reasoning	no time to answer	S. Literacy
3B	S. Reasoning	S. Reasoning	S. Reasoning	no time to answer	S. Reasoning
4B	S. Reasoning	S. Reasoning	S. Reasoning	no time to answer	S. Reasoning
Measure of variability					
1C	S. Literacy	S. Literacy	S. Literacy	no time to answer	S. Literacy
2C	S. Literacy	S. Reasoning	S. Literacy	no time to answer	S. Literacy
3C	S. Reasoning	S. Literacy	S. Literacy	no time to answer	S. Reasoning
4C	S. Reasoning	?	S. Literacy	no time to answer	S. Reasoning
Study design					
1D	S. Literacy	S. Literacy	S. Literacy	no time to answer	S. Literacy
2D	S. Literacy	S. Literacy	S. Literacy	no time to answer	S. Literacy
3D	S. Literacy	S. Literacy	S. Literacy	no time to answer	S. Literacy
4D	S. Literacy	S. Literacy	S. Literacy	no time to answer	S. Literacy
5D	S. Reasoning	S. Literacy	S. Literacy	no time to answer	S. Literacy
6D	S. Reasoning	S. Literacy	S. Reasoning	no time to answer	S. Reasoning
7D	S. Reasoning	S. Literacy	S. Literacy	no time to answer	S. Literacy
8D	S. Reasoning	?	S. Literacy	no time to answer	?
Hypothesis testing and p-values					
1F	S. Literacy	S. Reasoning	S. Literacy	S. Literacy	S. Literacy
2F	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Literacy
3F	S. Literacy	S. Literacy	S. Reasoning	S. Literacy	S. Reasoning
4F	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Literacy
5F	S. Literacy	S. Literacy	S. Reasoning	S. Reasoning	S. Literacy
9F	S. Literacy	S. Reasoning	S. Literacy	S. Literacy	S. Literacy
11F	S. Literacy	?	S. Literacy	S. Literacy	S. Literacy
12F	S. Literacy	S. Reasoning	S. Literacy	?	S. Literacy
6F	S. Reasoning	S. Reasoning	S. Reasoning	S. Reasoning	S. Literacy
7F	S. Reasoning	S. Reasoning	S. Literacy	S. Literacy	S. Literacy
8F	S. Reasoning	?	S. Literacy	S. Reasoning	S. Literacy
10F	S. Reasoning	S. Reasoning	S. Reasoning	S. Literacy	S. Literacy

Item	Author Classification	Expert A	Expert B	Expert C	Expert D
13F	S. Reasoning	S. Reasoning	S. Literacy	S. Reasoning	S. Reasoning
14F	S. Reasoning	S. Literacy	S. Literacy	S. Reasoning	S. Reasoning
Confidence interval					
1E	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Reasoning
2E	S. Literacy	S. Literacy	S. Reasoning	S. Literacy	S. Literacy
3E	S. Reasoning	S. Reasoning	?	S. Literacy	S. Reasoning
4E	S. Reasoning	?	S. Literacy	S. Literacy	S. Reasoning
Probability					
2G	S. Literacy	S. Reasoning	S. Literacy	no time to answer	S. Reasoning
3G	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Literacy
5G	S. Literacy	S. Literacy	S. Literacy	S. Reasoning	S. Literacy
1G	S. Reasoning	S. Literacy	S. Reasoning	S. Literacy	S. Reasoning
4G	S. Reasoning	S. Literacy	?	S. Literacy	S. Literacy
6G	S. Reasoning	S. Reasoning	S. Reasoning	S. Literacy	S. Reasoning
Bivariate data					
1H	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Literacy
2H	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Literacy
4H	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Reasoning
6H	S. Literacy	S. Literacy	S. Literacy	S. Literacy	S. Literacy
3H	S. Reasoning	S. Reasoning	S. Literacy	S. Reasoning	S. Reasoning
5H	S. Reasoning	S. Reasoning	S. Reasoning	?	S. Literacy

Appendix E: Think-aloud Interviews Correspondence

E1 - Email to instructors

Dear X,

I am writing to see if you might forward a message from me to your students who have completed EPSY 5261 during Summer 2015. I am developing an instrument for my dissertation research: the Reasoning and Literacy (REALI) instrument. The purpose of developing this instrument is to better understand the learning goals of statistical literacy and statistical reasoning. This instrument includes items on representations of data, measures of center and variability, study design, hypothesis testing and p-value, probability, and bivariate data. I am working with my advisers, Joan Garfield and Andrew Zieffler.

I am looking for 3 volunteers who would be willing to be interviewed as I ask them to read and respond to the questions on the assessment. During the interview, the students will be asked to verbalize everything that they are thinking as they work through the REALI instrument. This includes reading directions, explaining solutions to questions, and describing things they find confusing. This type of interview is a very important part of the instrument development, because will help expose flaws in the REALI instrument that should be rewritten or redesigned as well as provides insight that will inform scoring considerations for future use of the instrument.

Students that volunteer and are selected will be compensated with a $20 gift card if they complete the interview. The interview will consist of the following:
- The student and I will agree to meet at a mutually convenient time and location or via internet (using Skype and Google hangout).
- I will read a script describing the interview.
- The student completes the REALI instrument while describing their thinking.
- The student will be thanked and presented with (or emailed) the gift card.

Your students' feedback will be invaluable as I work toward creating a final version of this assessment. Would you be willing to copy and paste the text below the Student Email section into the body of an email to your class? Please add me as BCC.

Thank you for helping me with my research!

Sincerely,

Anelise G Sabbag
Doctoral Candidate in Quantitative Methods in Education
Department of Educational Psychology
University of Minnesota

STUDENT EMAIL - [Please copy & paste text below into an email to students and please use this subject line:

EMAIL SUBJECT LINE: Opportunity to help in statistics education research and win $20 gift card.

TO: Statistics Students
FROM: Anelise G. Sabbag, University of Minnesota

Hello,

I'm a PhD student at the University of Minnesota, and I'm working on designing an assessment to measure key learning outcomes of the introductory statistics course as a part of my dissertation research. I need feedback from students to help me improve the assessment, so I've asked your instructor to share this message with you.

I am looking for 3 volunteers to meet with me for about an hour at a mutually convenient time/place or via internet (Skype or Google hangout) to do something called a "think-aloud interview." Basically, you would just work through the assessment and talk with me about how you would answer the questions and describe things that are unclear or confusing. You will be audio-taped to produce a record of your responses for later analysis. Students that complete the interview will be given a $20 gift card for their participation.

If you're interested, please send me an email (sabb0013@umn.edu) and I will give you more information. I may not be able to include everyone that volunteers, so please contact me soon if you would like to participate!

Thanks!

Anelise G Sabbag
Doctoral Candidate in Quantitative Methods in Education
Department of Educational Psychology
University of Minnesota
sabb0013@umn.edu

E2 - Email for class visit and recruitment of students

Dear **X**,

I am writing to see if I might come to your STAT 3022 class to invite your students to participate in a cognitive interview as part of my research. The in-class invitation will only take 5 minutes.

I am developing an instrument for my dissertation research: the *Reasoning and Literacy* (REALI) instrument. The purpose of developing this instrument is to better understand the learning goals of statistical literacy and statistical reasoning. This instrument includes items on representations of data, measures of center and variability, study design, hypothesis testing and p-value, probability, and bivariate data. I am working with my advisers, Joan Garfield and Andrew Zieffler.

I am looking for 3 volunteers who would be willing to be interviewed as I ask them to read and respond to the questions on the assessment. During the interview, the students will be asked to verbalize everything that they are thinking as they work through the REALI instrument. This includes reading directions, explaining solutions to questions, and describing things they find confusing. This type of interview is a very important part of the instrument development, because will help expose flaws in the REALI instrument that should be rewritten or redesigned as well as provides insight that will inform scoring considerations for future use of the instrument.

Students that volunteer and are selected will be compensated with a **$20 gift card** after they complete the interview.

Your students' feedback will be invaluable as I work toward creating a final version of this assessment. **Would you allow me to go to your class tomorrow (11/12/2015) or next Monday (11/16/2015) to ask for volunteers?**

If it is not possible for me to come to your class, would you be willing to forward an invitation from me to your students asking for volunteers?

Thank you for helping me with my research!

Sincerely,

Anelise G Sabbag
Doctoral Candidate in Quantitative Methods in Education
Department of Educational Psychology
University of Minnesota

E3 - In class invitation script

Hello students.

I'm a PhD student at the University of Minnesota, and as part of my dissertation I'm working on designing an assessment that measures important learning outcomes of introductory statistics courses. I need feedback from some of you to help me improve the assessment, so I've asked your instructor if I could come to your class and make an invitation for you.
I am inviting you to participate in a one-hour interview that is designed to gain an understanding of your thinking strategies when you solve statistical questions from my assessment. You will be asked to talk aloud as you solve the questions. So you would just work through the assessment and talk with me about how you would answer the questions and describe things that are unclear or confusing. You will be audio-taped to produce a record of your responses for later analysis.
As an incentive to participate in this study, you will receive a **$20 gift card** if you complete the interview.
I am planning to conduct the interviews from **XX [date]** to **XX [date]**. If you are interested in participating please sign the sign-up sheet and provide your email address. I will then send you an email to let you know if you are selected to participate in the study as well as to set up a meeting time that is convenient for you.

You will be notified by **XX [date]** if you are selected to participate in the study, and you will be sent a survey to narrow down the times that you are available.
Thank you.

In-class Sign-up Sheet

Please fill in your information below if you are interested in participating in an interview with Anelise G Sabbag.

Name	Email address

E4 - Consent Form for think-aloud interview

CONSENT FORM FOR COGNITIVE INTERVIEW STUDENTS

The *Reasoning and Literacy Instrument* (REALI) is part of a research project which aims to measure statistical knowledge learning outcomes of the introductory statistics course. Please read this form carefully and ask any questions you have before agreeing to be in the study.

Background Information:

This study is being conducted by Anelise G. Sabbag, a doctoral candidate in the Department of Educational Psychology at the University of Minnesota, under the supervision of Dr. Joan Garfield and Dr. Andrew Zieffler.

The primary goal of the study is to develop an assessment tool to measure *statistical literacy* and *statistical reasoning* in students after they have completed (or nearly completed) an introductory statistics course. The information gathered from this assessment will used to better understand the learning goals of statistical literacy and statistical reasoning.

Procedures:
You will participate in a one hour interview that is designed to gain an understanding of what thinking strategies you use for the questions in the REALI assessment.

Each interview will be audio-taped to produce a record of your responses for later analysis. Excerpts of your interview may be used in research presentations or publications as an illustration of students' statistical literacy or statistical reasoning skills. These excerpts may be in the form of a transcription of your statements during the interview.

We are asking for your consent to do two things. First, we ask for your consent to audio-tape and record the interview. Second, we ask for your consent to include excerpts of your statements during the interviews in research presentations and publications.

Compensation:
You will be compensated with a $20 gift card for completing the think-aloud interview.

Risks and Benefits of Being in the Study:
As with all research, there is a chance that confidentiality could be compromised. However, we are taking appropriate steps to minimize this risk.

The benefit of participating is the opportunity to develop a better understanding of statistics, and of your own statistical literacy and reasoning capabilities.

Confidentiality:
The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify you as a participant. All research records will be stored on a secure server; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study:
Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with your institution or the course. If you decide to participate, you are free to withdraw at any time without affecting those relationships.

Contacts and Questions:

The researcher conducting this study is Anelise G. Sabbag under the supervision of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Andrew Zieffler, Ph.D. (Educational Psychology – Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Anelise G. Sabbag (sabb0013@umn.edu). You may also contact my academic advisors, Dr. Joan Garfield (jbg@umn.edu) and Dr. Andrew Zieffler (zief0002@umn.edu).

If you have any questions or concerns regarding the study and would like to talk to someone other than the researchers, you are encouraged to contact the Research Subjects' Advocate line, D528 Mayo, 420 Delaware Street S.E., Minneapolis, Minnesota 55455; telephone 612-625-1650.

Statement of Consent

I have read the above information. I have had the opportunity to ask questions and receive answers.
Please sign and return this consent form if you agree to let us use your responses in the research study described above. Please place an X next to each item below for which you do give your permission.

☐	I give permission to be recorded and audio-taped.
☐	I give permission to include excerpts of my statements in research presentations and publications.

Your Name (Please PRINT): _____

Signature: _____ Date: _____

E5 - Interview Protocol

Read to participant:
Thanks for meeting with me. Let me tell you a little more about what we'll be doing today.

1. I am testing a new statistics exam with the help of students such as yourself.
2. I'll give you the exam questions and you answer them, just like a regular exam.
3. However, my goal here is to get a better idea of how the questions are working. So I'd like you to *think aloud* as you answer the questions—just tell me *everything* you are thinking about as you go about answering them.
4. Please read the exam questions aloud while you are taking the exam.
5. Please keep in mind that I really want to hear all of your opinions and reactions. Do not hesitate to speak up whenever something seems unclear or is hard to answer.
6. Sometimes I will remind you to think aloud as you answer a question.
7. Finally, we'll do this for an hour, unless I run out of things to ask you before then.
8. Please take the time to look over the consent form and sign it at the bottom.

9. Do you have any questions before we start?
10. Now let’s take a look at the questions we are testing.

Think-Aloud Practice:

- Let's begin with a couple of practice questions. Remember to try to think aloud as you answer.
- Practice question 1: How many windows are there in the house or apartment where you live?
- [Probe as necessary]: How did you come up with that answer?
- Practice question 2: How difficult was it for you to get here to do the interview today: very difficult, somewhat difficult, a little difficult, or not at all difficult?
- [Probe as necessary]: Tell me more about that. Why do you say [ANSWER]?
- OK, now let's turn to the questions that we're testing.

Think-Aloud Interview:

- The student will be provided with the copy of the assessment.
- The student will be asked to complete the assessment while thinking aloud.
- Probes will be used if the student forgets to think aloud. Probes will not be used to elicit an answer from the student. Example probes include
    - “What are you thinking?”
    - “Keep talking”
    - If asked what something means ask “What do you think it means?”
- After the student completes the assessment, the student will be thanked, presented with the gift card compensation, and permitted to leave.

Appendix F: Pilot test Correspondence

F1 - Emails to instructors

Initial email:
Dear Stat Chat instructors,

I'm currently developing an assessment tool designed to measure important statistical learning outcomes (statistical literacy and statistical reasoning) of introductory statistics courses.

If you are a post-secondary introductory statistics instructor, I'm writing to ask for your help to pilot the *Reasoning and Literacy* (REALI) instrument. The REALI instrument includes items on representations of data, measures of center and variability, study design, hypothesis testing and p-value, probability, and bivariate data. The REALI instrument is composed of 48 multiple-choice items and I expect most students to take about 50 to 60 minutes to complete the test.

I am developing this instrument for my dissertation research at the University of Minnesota, where I am a doctoral candidate in the Department of Educational Psychology with a concentration in Statistics Education working under the supervision of my advisers, Joan Garfield and Andrew Zieffler.

I am requesting your help at this time to pilot the instrument with your students. The data gathered will be used to evaluate the validity and reliability of the REALI instrument.

The REALI instrument is designed to accommodate a wide variety of introductory statistics curricula. If you agree to participate, I will provide summary data that describes the results for your students as well as a summary of the results collected from other institutions for comparison upon request. You can administer the REALI instrument to your students as you see fit, and students may use any materials they wish though I ask that they work independently. You may want to provide credit or extra credit to your students as an incentive. You will receive your students' scores after they complete the exam.

The REALI instrument will be available **online** by December 3rd. Ideally, I would like you to administer the instrument to your students in the month of December. If this will not be possible, please let me know. I sincerely hope that you will be able to help during this phase of my study, and I thank you for your consideration.

If you are interested, please contact me at *sabb0013@umn.edu* with the following information:

1. Number of students in your class;

2. Institution name;

3. Short description of the curriculum (normal/t-distribution methods; resampling/simulation methods; etc).

Sincerely,
Anelise G. Sabbag
Doctoral Candidate in Quantitative Methods in Education
Department of Educational Psychology
University of Minnesota

2nd email:
Dear X,

Thank you for agreeing to administer my assessment tool to your students. I will soon email you the complete information to send to your students. But one important thing to note is that your students will be asked to enter a CODE in the beginning of the test. This code will be used to identify each one of your classes. Without this code I will not be able to give you a summary of the results for each class.

In a few moments I will send you 4 emails with the necessary information to send to each of your 4 classes. The reason I am sending 4 different emails is because you have 4 different classes and each class will have a **different code**. The only thing I ask you to do is to please forward each email to the correct class.

I will send an email to you on **December 19th** with the *names* and *total score* of students from your class who completed the assessment. I will send you the report we talked about by **December 23rd**.

Please note the test will be available online from **December 3rd** to **December 18th**.

If you have any concerns please let me know.

Thank you again.
Sincerely,
Anelise G. Sabbag

3rd email:
Dear X,

Please forward the text below to your students who are enrolled in the **MIS264A-PM** class.

Student Email
TO: Students
FROM: Anelise G Sabbag, University of Minnesota

Hello,

You are being asked to take a test designed to measure key learning outcomes of the introductory statistics course. The test consists of 48 multiple-choice questions and will take 50 to 75 minutes

to complete. You can use any resources that you want when completing the assessment *except other people*.

We would like to use your responses (which will be anonymized after providing your names to your instructor) as part of a research project. Having your data as part of this project is completely optional. You will be asked whether or not you want to participate in the research at the beginning of the test.

To complete the assessment and receive the extra credit, please click on the following link or copy and paste it into a web browser:

X

IMPORTANT:
One of the questions in the beginning of the test will ask you for a CODE. Your code is: X

The due date for completing the test is **December 18th, 2015 at 11:59 pm**.

If you have any questions about the test or trouble accessing the link, please email me at sabb0013@umn.edu.

Thank you!
Anelise G. Sabbag
Doctoral Candidate in Statistics Education
Department of Educational Psychology
University of Minnesota
sabb0013@umn.edu

F2 - Consent form

CONSENT FORM FOR PILOT TESTING

The Reasoning and *Li*teracy *I*nstrument (REALI) is part of a research project which aims to measure statistical knowledge learning outcomes of the introductory statistics course. Please read this form carefully and ask any questions you have before agreeing to be in the study.

Background Information:
This study is being conducted by Anelise G. Sabbag, a doctoral candidate in the Department of Educational Psychology at the University of Minnesota, under the supervision of Dr. Joan Garfield and Dr. Andrew Zieffler.

The primary goal of the study is to develop an assessment tool to measure *statistical literacy* and *statistical reasoning* in students after they have completed (or nearly completed) an introductory statistics course. The information gathered from this assessment will used to better understand the learning goals of statistical literacy and statistical reasoning.

Procedures:
You will complete an online version of the REALI instrument, which includes 40 items. It will take most students about 60 to 75 minutes to complete.

Compensation:
There is no compensation for participating in this research study.

Risks and Benefits of Being in the Study:

As with all research, there is a chance that confidentiality could be compromised. However, we are taking appropriate steps to minimize this risk.

The benefit of participating is the opportunity to develop a better understanding of statistics, and of your own statistical literacy and reasoning capabilities. The instructors of students participating in this study will be provided with the scores of their students.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify you as a participant. All research records will be stored on a secure server; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with your institution or the course. If you decide to participate, you are free to withdraw at any time without affecting those relationships.

Contacts and Questions:

The researcher conducting this study is Anelise G. Sabbag under the supervision of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Andrew Zieffler, Ph.D. (Educational Psychology – Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Anelise G. Sabbag (sabb0013@umn.edu). You may also contact my academic advisors, Dr. Joan Garfield (jbg@umn.edu) and Dr. Andrew Zieffler (zief0002@umn.edu).

If you have any questions or concerns regarding the study and would like to talk to someone other than the researchers, you are encouraged to contact the Research Subjects' Advocate line, D528 Mayo, 420 Delaware Street S.E., Minneapolis, Minnesota 55455; telephone 612-625-1650.

Statement of Consent

Please click the circle below if you agree to participate in this research study.

	I have read the above information and I give permission for my responses to assessment items to be included in any analyses, reports, or research presentations made as a part of this research project.

***Online Test Instructions**

You will now start the REALI online test. This test includes 40 questions. Please read each question carefully and chose one of the options. For all open-ended questions, please describe your reasoning. You can click the next button to go to the next question. You can also go back to previous question(s) to review or change your answer(s) by clicking the back button.

Appendix G: Field test Correspondence

G1 - Emails to instructors

Initial Email:
Dear Instructor,

I am writing to request your assistance in helping me recruit introductory statistics students (at either the undergraduate or graduate level) to take part in the field-testing of an assessment tool designed to measure important statistical literacy and statistical reasoning learning outcomes of introductory statistics courses. This field-testing is part of my dissertation research, under the supervision of my advisers, Joan Garfield and Andrew Zieffler at the University of Minnesota.

The assessment is composed of 40 multiple-choice items assessing topics such as representations of data; measures of center and variability; study design; hypothesis testing and *p*-value; probability; and bivariate data and accommodates a wide variety of introductory statistics curricula. Students should be able to complete the assessment in less than an hour.

The instrument is administered online. Other than asking that students work independently to complete the assessment, there are no other constraints on the administration (e.g., it could be completed in-class or outside of class). To increase student participation and effort when completing the assessment, you may want to provide credit or extra credit to your students. To incentivize you to recruit students, if your students complete the assessment, I will provide you with summary data describing your students' performance. I will also provide a pooled summary of performance across all participating institutions.

Ideally, I would like you to administer the instrument to your students by **May 2nd**. If this will not be possible, please let me know, and I can work with you to find a time that works. Students should have the opportunity to learn the concepts before taking the assessment, so I ask that you please administer the assessment close to end of the course.

If you are interested, please contact me at sabb0013@umn.edu with the following information:

- Institution name;
- Course name;
- Number of sections;
- Number of students in each section;
- Short description of the curriculum (normal/t-distribution methods; resampling/simulation methods; etc).

I sincerely hope that you will be able to help during this phase of my study, and I thank you for your consideration.

Sincerely,
Anelise G. Sabbag
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

2nd email

Dear X,

Thank you again for helping me with my research!
Below is the email with the instructions that your students need to complete the assessment.
Please forward the text below to your students.

Thank you very much,
Anelise

Student Email

TO: Students
FROM: Anelise G Sabbag, University of Minnesota

Hello,

You are being asked to take a test designed to measure key learning outcomes of the introductory statistics course. The test consists of 40 multiple-choice questions and will take 50 to 75 minutes to complete. You can use any resources that you want when completing the assessment *except other people*.

We would like to use your responses (which will be anonymized after providing your names to your instructor) as part of a research project. Having your data as part of this project is completely optional. You will be asked whether or not you want to participate in the research at the beginning of the test.

To complete the assessment please click on the following link or copy and paste it into a web browser:

X

IMPORTANT:
One of the questions in the beginning of the test will ask you for a CODE.
Your code is: X

If you have any questions about the test or trouble accessing the link, please email me at sabb0013@umn.edu.

Thank you!
Anelise G. Sabbag
Doctoral Candidate in Statistics Education
Department of Educational Psychology
University of Minnesota
sabb0013@umn.edu

G2 - Consent form

CONSENT FORM FOR FIELD TESTING

Please read this consent statement and indicate whether you agree to participate in this study.

Background Information:
This study is being conducted by Anelise G. Sabbag, a doctoral candidate in the Department of Educational Psychology at the University of Minnesota, under the supervision of Dr. Joan Garfield and Dr. Andrew Zieffler.
The primary goal of the study is to develop an assessment tool to measure statistical literacy and statistical reasoning in students after they have completed (or nearly completed) an introductory statistics course. The information gathered from this assessment will used to better understand the learning goals of statistical literacy and statistical reasoning.

Procedures:
You will complete an online version of the REALI instrument, which includes 40 items. It will take most students about 50 to 75 minutes to complete.

Risks and Benefits of Being in the Study:
As with all research, there is a chance that confidentiality could be compromised. However, the research is taking appropriate steps to minimize this risk.
The benefit of participating is the opportunity to develop a better understanding of statistics, and of your own statistical literacy and reasoning capabilities. The instructors of students participating in this study will be provided with the scores of their students.

Confidentiality:
The records of this study will be kept private. In any sort of report the researcher might publish, the researcher will not include any information that will make it possible to identify you as a participant. All research records will be stored on a secure server; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study:
Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with your institution or the course. If you decide to participate, you are free to withdraw at any time without affecting those relationships.

Contacts and Questions:
The researcher conducting this study is Anelise G. Sabbag under the supervision of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Andrew Zieffler, Ph.D. (Educational Psychology – Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Anelise G. Sabbag (sabb0013@umn.edu). You may also contact my academic advisors, Dr. Joan Garfield (jbg@umn.edu) and Dr. Andrew Zieffler (zief0002@umn.edu).
If you have any questions or concerns regarding the study and would like to talk to someone other than the researchers, you are encouraged to contact the Research Subjects' Advocate line, D528 Mayo, 420 Delaware Street S.E., Minneapolis, Minnesota 55455; telephone 612-625-1650.

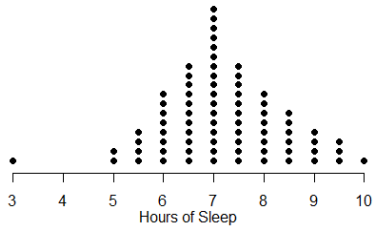
STATEMENT OF CONSENT
I have read the above information and I give permission for my responses to be included in any analyses, reports, or research presentations made as a part of this research project.

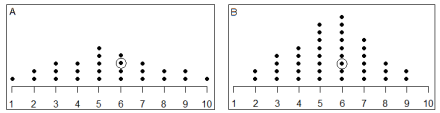
Appendix H: Haberman Analysis for the Raw Subscores

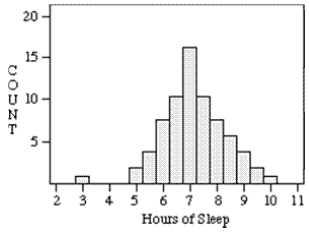
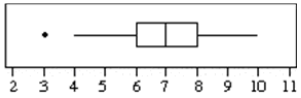
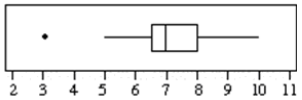
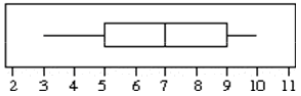
Model	Learning Goal	PRMSE(S_z) Total Score	PRMSE(S_x) Subscore	Difference
Raw Scores	Literacy	0.8835	0.7620	-0.1215
	Reasoning	0.8841	0.7816	-0.1025

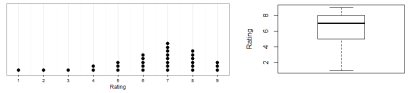
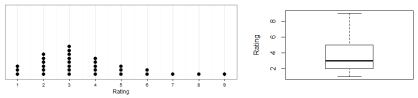
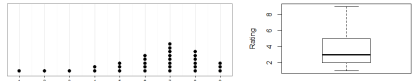
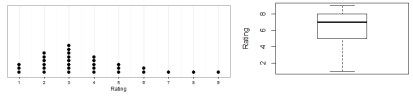
Appendix I: Item Changes

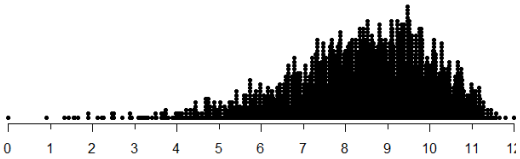
I1 - Minor changes done to items based on expert review and think-aloud interviews.

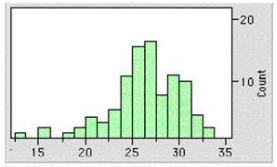
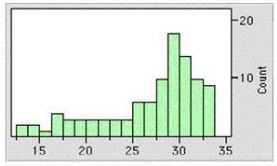
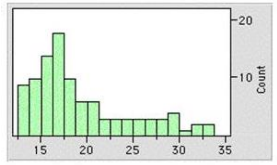
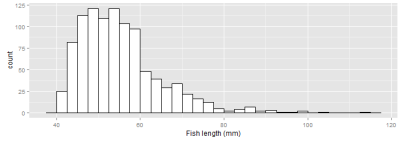
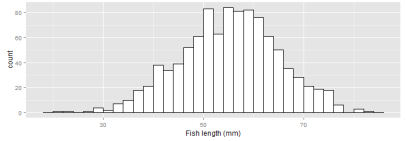
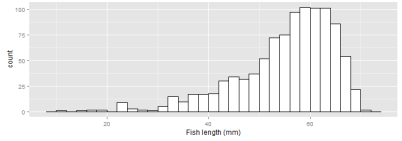
#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
Representations of Data				
1A	The following graph shows the distribution of hours slept the previous night by a group of college students.  Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable. a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five. b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1. c. Many students seem to be getting 7 hours of sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep. d. The distribution of hours of sleep is somewhat normal, with an outlier at 3. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.	One expert commented that alternative D was the most logical choice because it had language usually used in a statistics textbook. In addition, this expert also suggested moving the word "statistically" to the end of the sentence in the stem. This item was originally from the BLIS instrument (Ziegler, 2014). Based on the field test results reported by Ziegler, option D (correct answer) was the most attractive option (75% of the students chose this option) and option B was the least attractive (7.3%). As an attempt to shorten this item and improve its characteristics, option B was deleted. The wording of the remaining three options was also changed to make them less distinct in terms of the words used to describe the distributions. The phrase "The distribution is somewhat bell-shaped" was added to option A. Option C was changed to "Many students seem to be getting 7 hours of sleep at night, but some students slept more and some slept less. The minimum value is 3 and the maximum is 10." Option D was changed to "the distribution of hours of sleep is somewhat bell-shaped. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour."	While reading through this item, some students recognized that the alternatives were not balanced because only alternative C provided the standard deviation as part of the description of the distribution. Therefore, it was decided to include the standard deviation in the description of one more alternative (option A). Out of the four students who did the pilot-test, two of them answered item 1A incorrectly. After additional review of this item and conversations with statistics educators from the University of Minnesota, it was determined that adding a more specific context to the item would encourage students to pay attention to the plausibility of each of the interpretations. The stem of this item changed to: "A high school teacher is concerned with whether her students are getting enough sleep. She asked each one of her students to report the number of hours slept the previous night. A plot of the results is given below. She would like to report her findings to the school principal using appropriate statistical language to describe and interpret the distribution in context. Select the **most appropriate** statement that she could use to summarize the results."	

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
2A	Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights for 39 random samples of 3 pebbles each, with all mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.  d. No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams. e. Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6. f. Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B represents the average weight of 3 pebbles.	Based on expert feedback, the word "average" was changed to the word "mean" to be consistent with the wording used in the stem of the item.		The word "figure" was added to the stem of the item before the letters "A" and "B" to be consistent with the wording in the alternative options.

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
4A	The following graph shows a distribution of hours slept last night by a group of college students.  Which box plot seems to be graphing the same data as the histogram above? a.  b.  c.  			Grammar was fixed by changing the word "box plot" to "boxplot".

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
5A	One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. After analyzing the answers from the students, the instructor interpret the data saying "**A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding."** The instructor asked two of his students to create a graphical representation of the data, based on his interpretation above. Beth created a dotplot and Allan created a boxplot. Which dotplot/boxplot pair better aligns with the description given by the instructor? a.  b.  c.  d. 	Based on expert feedback, a grammar mistake was fixed. The word "interpret" was changed to "interprets".		The stem of Item 5A was changed to increase clarity: An instructor gave his introductory statistics students a survey. One of the questions read, "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. After analyzing the answers from the students, the instructor interpreted the data as follows: "**A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding."** The instructor asked two of his students to create a graphical representation of the data, based on his interpretation above. Beth created a dotplot and Allan created a boxplot. Which dotplot/boxplot pair best aligns with the description given by the instructor?

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
Measures of Center				
1B	According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the mean? b. For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700. c. For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700. d. For all dog owners in this sample, about half were above $1,700 and about half were below $1,700. e. For most owners, the first-year costs for owning a large-sized dog is $1,700.	Based on expert feedback, the word “average” was changed to the word “mean”.		
2B	Which of the following intervals is MOST likely to include the mean of the distribution below?  a. 6 to 7 b. 8 to 9 c. 9 to 10 d. 10 to 11	Originally, this item was from BLIS and the descriptive data analysis reported in Ziegler (2014) showed that only 1% of the students chose alternative D. Therefore, this alternative seemed to not be a plausible distractor and was deleted.		

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
4B	A study examined the length of a certain species of fish from one lake. The plan was to take a random sample of 100 fish and examine the results. The mean length was 26.8mm, the median was 29.4mm, and the standard deviation was 5.0mm. Which of the following histograms is most likely to be the one for these data? a.  b.  c.  		When solving Item 4B, all students presented difficulties differentiating between alternatives A and B. All stated that both alternatives were skewed to the left but it was very hard to choose between them. Therefore, this item was changed to present three alternative options that would be significantly different from each other. In addition, it was also noticed that none of the students used the information about the standard deviation presented in the stem of the item. Therefore, this extraneous information was deleted and the stem was also shortened. The plots for options A, B, and C are    	

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
Measures of Variability				
2C	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 to be -2.30. What do we know? e. The standard deviation was calculated incorrectly. f. Most students received negative scores. g. Most students scored below the mean. h. None of the above.			The stem of Item 2C was changed to increase clarity: A teacher gives a science test with 15 questions. For each question, 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 to be -2.30. What do we know?

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
3C	Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following data for 5 days of travel on each route. Country Route - 17, 15, 17, 16, 18 City Route - 18, 13, 20, 10, 16 It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose? a. The Country Route, because the times are consistently between 15 and 18 minutes. b. The City Route, because she can get there in 10 minutes on a good day and the average time is less than for the Country Route. c. Because the times on the two routes have so much overlap, neither route is better than the other. She might as well flip a coin.	Based on expert feedback, the sentence "on the time of travel in minutes" was added to the stem of the question.	See Section 4.2 for detailed description of changes for this item. Changes were made to this item to show more evidence of statistical reasoning. The first change done to the item was to present a graphical representation of the data instead of presenting the raw numbers. In addition, the explanations shown in each of the alternatives were deleted to increase item difficulty.	

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
Study Design				
1D	The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population. d. The population is all American adults in 2013. The sample is the 21% of American adults that have had an email or social networking account compromised. e. The population is the 1,002 American adults surveyed. The sample is all American adults in 2013. f. The population is all American adults in 2013. The sample is the 1,002 American adults surveyed.			The last sentence of Item 1D was changed to increase clarity: Identify the population about which the Pew Research Center can make inferences from the survey results. Identify also the sample from that population.
3D	A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level? d. Observational study e. Randomized experiment f. Survey			The option C of Item 3D was changed to follow the same format as option A. So the word "study" was added to option C In this way all alternatives contain 2 words and are therefore balanced.

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
4D	A research study randomly assigned participants into two groups. One group was given Vitamin E to take daily. The other group received only a placebo pill. The research study followed the participants for eight years. After the eight years, the proportion of each group that developed a particular type of cancer was compared. What is the primary reason that the study used random assignment? a. To ensure that the groups are similar in all respects except for the level of Vitamin E. b. To ensure that a person doesn't know whether or not they are getting the placebo. c. To ensure that the study participants are representative of the larger population.		This item assessed students' ability to reason about the primary purpose of random assignment. Three of the four students who did the think-aloud interview choose alternative B for this item because they were thinking about the placebo effect. Therefore, they did not present clear understanding about the primary purpose of random assignment. In addition, these students did not pay attention to the word "primary" in the stem of the question. This is an important word to distinguish between alternatives A and B. Therefore, to call more attention toward this word, it was decided to bold it. One of the students – this student was an international student – was also unclear about the wording of alternative C. The student mentioned that there was no level of the vitamin and that a person would either take the vitamin or not. Therefore, the wording of alternative A was changed to not include the word "level": to ensure that the groups are comparable except for the treatment variable.	

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
5D	A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election. Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why. a. Valid, because the sample size is large enough to represent the population. b. Valid, because 67% is far enough above 50% to predict a majority vote. c. Invalid, because the sample is too small given the size of the population. d. Invalid, because the sample may not be representative of the population.	Based on expert feedback, the wording on option C was changed to increase difficulty. The new wording of option C was “invalid, because the sample is only 1/50th of the size of the population”.	See Section 4.2 for detailed description of changes for this item. Because students were not even reading options A and B, these alternatives were deleted and a new invalid alternative was created: invalid, because not all viewers had the opportunity to respond to the poll.	

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
7D	A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method? a. This is a simple random sample. It will give an accurate estimate. b. Because the sample is so small, it will not give an accurate estimate. c. Because all fans had a chance to be asked, it will give an accurate estimate. d. The sampling method is biased. It will not give an accurate estimate.	One expert suggested changing option B to address a common student's misconception that larger samples gives an accurate estimate despite study design. The new wording of option B was "because the sample is large, it will give an accurate estimate." In addition, to align the stem of the question to the new option B, the sample size in the stem was changed to 400.	See Section 4.2 for detailed description of changes for this item. To balance the alternatives, the sample size of the question was decreased to 250 and option B was changed to have a negative statement.	

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
Hypothesis Testing and *p*-values				
2F	A researcher wants to answer the following research question: Do sleep apnea patients look more attractive after receiving a new treatment (continuous positive airway pressure therapy)? Twenty patients with sleep apnea were given the new treatment. Pictures of the patients were taken before the treatment as well as two months after the treatment. An independent rater was shown the before and after pictures for each patient in a random order and asked to determine which picture the patient looked more attractive in. Seventy percent (70%) of the patients were rated as more attractive in the after picture. Is there a need to conduct a hypothesis test to determine whether the treatment is effective, or could the researcher just use the sample statistic (70%) as evidence of the effectiveness? d. The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%. e. The researcher should conduct a hypothesis test because a hypothesis test is always appropriate. f. The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.	One expert commented that options A and B were weak. This item was originally from the BLIS instrument (Ziegler, 2014). Based on the field test results reported by Ziegler, the percentages of students who chose option A, B, and C are respectively 9.7%, 17.4%, and 72.9%. Therefore, option A seemed to be indeed a weak response option. To improve option A, the sentence "therefore, the result did not happen by chance" was included.		

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
3F	The Amherst H. Wilder Foundation conducts surveys with homeless people in Minnesota. In their 2012 survey of 4,465 homeless adults, one question asked was "In the last 30 nights, how many nights have you spent outside, in a vehicle or vacant building, or some other place not intended for housing?" One research question that the surveyors had was "Is there a difference between males and females with regards to the average number of nights spent in a place not intended for housing?" Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question? e. There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing. f. There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing. g. There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing. h. There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.			The stem of Item 3F was changed to increase clarity. The word "had" was changes to "asked".

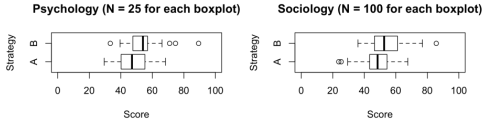
#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
6F	One hundred student-athletes attended a summer camp to train for a particular track race. All 100 student-athletes followed the same training program in preparation for an end-of-camp race. Fifty of the student-athletes were randomly assigned to additionally participate in a weight-training program along with their normal training (the training group). The other 50 student-athletes did not participate in the additional weight-training program (the non-training group). At the end of the summer camp, all 100 student-athletes ran the same race and their individual times (in seconds) were recorded. The mean speed of the training group was 44 seconds, and the mean speed of the non-training group was 66 seconds. The standard deviation for the non-training group was 20 seconds. Consider the following possible values for the standard deviation of the training group. Which of these values would produce the strongest evidence of a difference in means between the two groups? a. 10 seconds b. 20 seconds c. 30 seconds	Based on expert feedback, the words “in means” was added to the end of the stem.		

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
9F	A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes (p = 0.287). What does this mean? a. The distribution of the GPAs for male and female scholarship athletes are identical except for 28.7% of the athletes. b. The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287. c. There is a 28.7% chance that a pair of randomly chosen male and female scholarship athletes would have a significant difference assuming that there is no difference. d. There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between male and female scholarship athletes as that observed in the sample assuming that there is no difference.	Some experts pointed out that the correct alternative (option D) was longer that the other alternatives. Therefore, the wording of the alternative options B and D was changed. Option B became "the difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287 or larger, assuming that there is a difference between the means." Option D became "There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between males and females as that observed in the sample assuming that there is no difference."		

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
10F	Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached. The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean? a. The sample sizes were too small to detect a true difference between the coached and not-coached students. b. The observed difference between coached and not-coached students could occur just by chance alone even if coaching really has no effect. c. The increase in test scores makes no difference in getting into college since it is not statistically significant. d. The study was badly designed because they did not have equal numbers of coached and not-coached students.			The correct answer on Item 10F (alternative B) was longer than the other alternatives. Therefore, changes were made to the alternative B to decrease its length. In addition, the first alternative did not seem to be a plausible distractor because it was referring to the sample sizes of 4,200 and 500 as “too small”. Therefore, this first option was deleted.

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
11F	A researcher is interested if there is a significant difference in the average number of hours watching television between males and females 8th grade students in the US. After gathering and analyzing the data, the researcher found that the difference in the average number of hours between the two groups was 4.37 hours. The researcher conducted a test to verify if this difference in means was statistically significant and found a p-value of 0.001. What would be the correct conclusion the researcher needs to make? a. The difference between groups in the average number of hours watching television did happen by chance because the p-value is so small. b. The difference between groups in the average number of hours watching television is NOT statistically significant because the p-value is too small. c. There IS strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there is NO difference. d. There is NOT strong evidence of a difference in mean number of hours watching television between males and females in the population, but it is possible that in reality there IS a difference.	Based on expert feedback, it was noticed that the correct option (C) was the longest one. Therefore, the word "strong" was deleted from option C to make it the same length as option D. In addition, one of the experts suggested to add a rate (hours/week) to make the problem more authentic.		Options B, C, and D were changed so that negative words would be bold but not capitalized.

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
12F	A research article reports the results of a new drug test. The drug is hypothesized to decrease vision loss in people with macular degeneration more effectively than the current treatment. The article reports a *p*-value of 0.04 in the analysis section. Which option below presents the correct interpretations of this p-value? a. We conclude that the new drug is not effective because there is only a .04 probability that the drug is more effective than the current treatment. b. We conclude that the new drug is effective because results like they found, or results even more favorable to the new drug, would only happen 4% of the time if the drug was not effective. c. We conclude that the new drug is effective because there is only a 4% chance that it's not. d. We conclude that the new drug is not effective because the difference in the proportion of macular degeneration patients with vision loss between the two treatments is only 0.04.	Based on expert feedback, a grammar mistake was fixed. The word "results" was changed to "result". In addition, it was noticed by the experts that the correct option (B) was the longest one. Therefore, changes were done to option B to decrease its length. Option B became "We conclude that the new drug is effective because a result like they found, or more extreme, would only happen 4% of the time if the drug was not effective."		

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
13F	Two experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100). The two different experiments were conducted with students who were enrolled in two different subject areas: psychology and sociology. Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence <u>against</u> the claim, "neither strategy is better than the other"? Why?  a. Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment. b. Psychology, because there are more outliers in strategy B from the Psychology experiment, indicating that strategy B did not work well in that course. c. Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment. d. Sociology, because the sample size is larger in the Sociology experiment, which will produce a more accurate estimate of the difference between the two strategies.	One expert mentioned that negative statements can disadvantage some students. Therefore, the stem of the question was changed from "which experiment provides the stronger evidence <u>against</u> the claim" to "which experiment provides the stronger evidence supporting the claim."		For Item 13F, the correct alternative (option D) was longer that the other alternatives. Therefore, the wording of alternative D was changed to decrease its length. New option D was "Sociology, because the sample size is larger in the Sociology experiment. This will produce a more precise estimate of the difference between strategies."

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
Probability				
1G	Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 50 times and the other student flips a coin 100 times. Which student is more likely to get 48% to 52% of their coin flips heads up? a. The student who flips the coin 50 times because the percent that are heads up is less likely to be exactly 50%. b. The student who flips the coin 100 times because that student has more chances to get a coin flip that is heads up. c. The student who flips the coin 100 times because the more flips that are made will increase the chance of approaching a result of 50% heads up. d. Neither student is more likely because the flipping of the coin is random and therefore you cannot predict the outcome of the flips.	One expert suggested adding the word "fair" to the stem of the question to clarify that the probability of heads and tails are 0.5 and 0.5. In addition, the expert also suggested decreasing the length of option C (the correct option), so it would not be the longest. The new wording of option C was "the student who flips the coin 100 times because the more flips that are made will increase the likelihood of a result of 50% heads up". In addition, this item was originally from BLIS (Ziegler, 2014) and according to the field test results reported by Ziegler, option A was the least chosen option with around 6% of the students choosing it. Therefore, option A was deleted.	See Section 4.2 for detailed description of changes for this item. The number of coin flips was changed to 25 and 125 because one student was confused between the *number of flips* and the *probability of heads* in the long term.	

#	Item	Changes after the expert review	Changes after think-aloud interview	Additional changes
4G	A game company created a little plastic dog that can be tossed in the air. It can land either with all four feet on the ground, lying on its back, lying on its right side, or lying on its left side. However, the company does not know the probability of each of these outcomes. Which of the following methods is most appropriate to estimate the probability of each outcome? a. Since there are four possible outcomes, assign a probability of 1/4 to each outcome. b. Toss the plastic dog many times and see what percent of the time each outcome occurs. c. Simulate the data using a model that has four equally likely outcomes.	Based on expert feedback, a grammar mistake was fixed. The word “since” was changed to “because”.		

I2 - Major changes done to the items after the think-aloud interviews.

Measures of variability

To answer Item 3C (Figure I1a) correctly (behaviors needed), students needed to (1) know how to summarize and compare two datasets in terms of center and variation, (2) know that smaller variation means the data are more consistent, and (3) make connections between the mean and a measure of variation for each dataset. However, during the think-aloud interviews, only one student considered both the mean and the standard deviation of each route. The other two students only looked at the variation of the data but even so they were able to answer the question correctly. The fourth student did not answer this item.

Based on the three behaviors listed above, Item 3C was considered a statistical reasoning item. Students' answers, however, during the think-aloud interview showed that students were correctly answering the item without making connections between statistical concepts. Therefore, changes were made to this item to require more evidence of statistical reasoning. The first change done to the item was to present a graphical representation of the data instead of presenting the raw numbers. This was done with the intent to make students reason between a graphical representation of data and the measures of center and variation. In this way students were making connections between at least two statistical concepts which supported the assumption that this is a statistical reasoning item.

It is important to mention that this question was originally from the CAOS instrument and it had an acceptable but not good discrimination (0.20). In addition, this question was also considered very easy (percent correct was 80%). A similar version of this question was included in the GOALS instrument. This version of the item had poor discrimination (0.125) and also was very easy with a percent correct of 93.2%. Usually, if an item is very easy then almost all students

can correctly answer it; therefore, these items most often do not discriminate well between students with low and high ability. Therefore, it was decided to try to slightly increase the level of difficulty of this question. During the think-aloud interview, students payed close attention to the word "consistently" in the alternative option A and it seemed that this was a significant reason why students were choosing this option. Therefore, the explanations shown in each of the alternatives were deleted to increase item difficulty. The final version of Item 3C can be seen in Figure I1b.

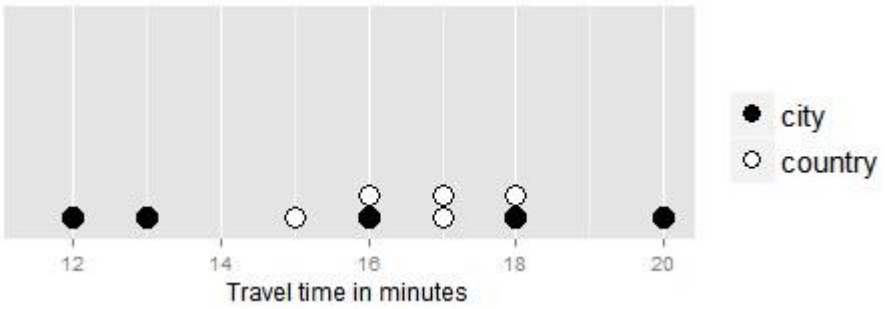
a) Used in the think-aloud interview:
Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following data on the time of travel in minutes for 5 days of travel on each route.

Country Route - 17, 15, 17, 16, 18
City Route - 18, 13, 20, 10, 16

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

a. The Country Route, because the times are consistently between 15 and 18 minutes.
b. The City Route, because she can get there in 10 minutes on a good day and the average time is less than for the Country Route.
c. Because the times on the two routes have so much overlap, neither route is better than the other. She might as well flip a coin.

b) Used in the Pilot test:
Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following commute times, in minutes, for 5 days of travel on each route.



It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

a. The City Route
b. The Country Route
c. Neither route is better than the other.

Figure I1. Version of Item 3C.

Study design

Item 5D (Figure I2a) was originally an item from GOALS designed to assess students' ability to reason about the factors that allow a sample of data to be representative of the population. However, in the expert review this item was categorized by three experts as a statistical literacy item. Because of the uncertainty regarding the classification of this item, it was decided to include it in the think-aloud interviews. In this way, it would be possible to verify if students were indeed making connections between the statistical concepts when answering Item 5D.

This item was originally categorized as a statistical reasoning item because its alternatives addressed a variety of statistical concepts. Therefore, by reading each alternative, the students would have to make connections and consider all the statistical concepts presented. However, three of the four students, during the think-aloud interviews, did not even read the first two alternatives. As soon as these students finished reading the stem of the item, they recognized that the statement was invalid, and therefore they only looked at alternatives C and D.

Based on how students were answering this item in the think-aloud interviews, it seemed that the item was not behaving as a statistical reasoning item. Therefore, changes were made to Item 5D. Because students were not even reading options A and B, these alternatives were deleted and a new *invalid* alternative was created. In this way, all options were invalid and this would force students to go through each of them and thus reason with different statistical concepts. The new version of Item 5D is shown in Figure I2b.

a) Used in the think-aloud interview:
A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.
Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

a. Valid, because the sample size is large enough to represent the population.
b. Valid, because 67% is far enough above 50% to predict a majority vote.
c. Invalid, because the sample is only 1/50th of the size of the population.
d. Invalid, because the sample may not be representative of the population.

b) Used in the Pilot test:
A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.
Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

a. Invalid, because the sample is only 1/50th of the size of the population.
b. Invalid, because not all viewers had the opportunity to respond to the poll.
c. Invalid, because the sample may not be representative of the population.

Figure I2. Versions of Item 5D.

Item 7D (Figure I3a) measured students' ability to recognize biased and unbiased sampling methods. In addition, students had to recognize that the sampling method used in the stem of the question was not a random sample. Finally, students also had to understand that the difference between accuracy was related to study design and not sample size. Because of these behaviors just listed, this item was initially categorized as a statistical reasoning item. However, in the expert review, this item had ratings from three experts and all of them categorized this as a statistical literacy item. However, it is important to notice that the reasoning part of this item is located in the four alternative options. Each alternative addresses a difference statistical concept, therefore, forcing students to reason with these concepts while answering the question. So this item was

included in the think-aloud interviews to gather evidence that students were indeed reasoning while answering it.

a) Used in the think-aloud interview:
A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 400 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?

a. This is a simple random sample. It will give an accurate estimate.
b. Because the sample is large, it will give an accurate estimate.
c. Because all fans had a chance to be asked, it will give an accurate estimate.
d. The sampling method is biased. It will not give an accurate estimate.

b) Used in the Pilot test:
A sportswriter wants to know the extent to which football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?

a. This is a simple random sample. It will give an accurate estimate.
b. Because the sample is so small, it will not give an accurate estimate.
c. Because all fans had a chance to be asked, it will give an accurate estimate.
d. The sampling method is biased. It will not give an accurate estimate.

Figure I3. Versions of Item 7D.

While answering Item 7D, all students went through each of the alternatives and thought carefully about the content of each alternative and the design of the study described in the stem. Therefore, students were indeed connecting the different statistical concepts presented in each of the alternatives, supporting then the assumption that this was indeed a statistical reasoning item.

This item was originally from the AIRS instrument, but changes were done to this item based on the expert review (see previous section). However, during the think-aloud interview one of the students noticed and mentioned that there were three affirmative options (A, B, and C) and only one negative option (D) so if he was short on time he would choose option D just because it was the only negative sentence. Because of this student's comment, the alternatives from Item 7D were changed to be more balanced (2 affirmative and 2 negative options). The sample size used in

this item was 400 and alternative B was referring to this sample size. To balance the alternatives, the sample size of the question was decreased to 250 and option B was changed to have a negative statement. These changes returned this item to its original form in the AIRS instrument (Figure I3b).

Item 8D (Figure I4) was originally from the AIRS instrument and it was designed to assess if students were able to evaluate the results of hypothesis testing considering some statistical concepts such as sample size, practical significance, effect size, data quality, and soundness of the method.

Two of the four students who did the think-aloud interview expressed concerns related to the content addressed by this item. Both students mentioned that this item was addressing two groups of patients: those with severe illness and those without severe illness. In addition, they mentioned that those patients with more severe illness would lead to lower survival times. Thus these two students recognized possible confounding variables that would have affected the study. However, they were not making the connection to study design and the lack of random assignment. In addition, one of the students mentioned she knew the statement was invalid, but she did not agree with any of the four response options presented.

No changes were made to Item 8D because it seemed that students' concerns were more related to a misconception they had, which is not understanding that confounding variables will be present when random assignment is not used in the study. However, after the pilot test, this item was re-evaluated to verify if it should remain in the instrument.

A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that 'prescription medications only' was the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?

a. Yes, because medication patients lived longer.
b. No, because doctors chose the treatments.
c. Yes, because the study was a comparative experiment.
d. No, because the patients volunteered to be studied.

Figure I4. Item 8D.

Probability

Item 1G (Figure I5a) was originally a BLIS item assessing students' ability to understand that randomness cannot be outguessed in the short term but patterns can be observed over the long term. In the expert review, half of the experts categorized this item as a statistical reasoning and half of the experts as a statistical literacy item. The author then re-evaluated the categorization of this item and considered that according to its behaviors and the working definitions, Item 1G could have been considered a statistical reasoning item. However, uncertainty remained regarding the categorization of this item. This was the reason why Item 1G was included in the think-aloud interview.

Two out of the four students could not answer this item because of the lack of time. The two students who answered this item both showed evidence that they were reasoning between the concepts of randomness, probability in the short run, and probability on the long run. For instance, both students' first thought that the probability of each side of the coin for a single flip was around 50%. Then they extended this idea of probability but to a longer sequence of flips recognizing that in the long run the number of heads and tails would even out. This kind of thinking process suggests that students were indeed reasoning with the statistical concepts.

One student while answering this question mentioned that 100 flips would be the correct answer because about half of 100 is closer to 50. Therefore, this student was confused regarding the *number of flips* and the *probability of heads* in the long term. Thus this item was changed to not use the numbers 100 and even 50. The new version of Item 1G is shown in Figure I5b.

a) Used in the think-aloud interview:
Two students are flipping fair coins and recording whether or not the coin landed heads up. One student flips a coin 50 times and the other student flips a coin 100 times. Which student is more likely to get 48% to 52% of their coin flips heads up?

a. The student who flips the coin 50 times because the less flips that are made will increase the likelihood of a result of 50% heads up.
b. The student who flips the coin 100 times because the more flips that are made will increase the likelihood of a result of 50% heads up.
c. Neither student is more likely because the flipping of the coin is random and therefore you cannot predict the outcome of the flips.

b) Used in the Pilot test:
Two students are flipping fair coins and recording whether or not the coin landed heads up. One student flips a coin 25 times and the other student flips a coin 125 times. Which student is more likely to get 48% to 52% of their coin flips heads up?

a. The student who flips the coin 25 times because the less flips that are made will increase the likelihood of a result of 50% heads up.
b. The student who flips the coin 125 times because the more flips that are made will increase the likelihood of a result of 50% heads up.
c. Neither student is more likely because the flipping of the coin is random and therefore you cannot predict the outcome of the flips.

Figure I5. Versions of Item 1G.

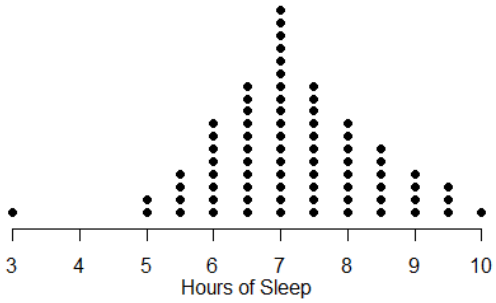
I3 - Major changes done to items after the pilot test.

Representations of Data.

There was a total of five items in this area: three statistical reasoning and two statistical literacy items. The aim for this area of learning was to have four total items (two statistical literacy and two statistical reasoning items). Therefore, there was a need to delete one statistical reasoning item. Item 4 was the statistical reasoning item that was deleted, because it was a badly discriminating item.

Based on the pilot data, Item 1/1A (Figure I1a) was also flagged as a badly discriminating item (item discrimination = 0.12 < 0.20). As mentioned in sections 4.1 and 4.2, this item was originally from BLIS, but modifications were done to this item based on feedback from the expert review and think-aloud interviews. However, these modifications led to worse item characteristics than the original item. Therefore, the alternative options of Item 1/1A were returned to the original format (Figure I1b).

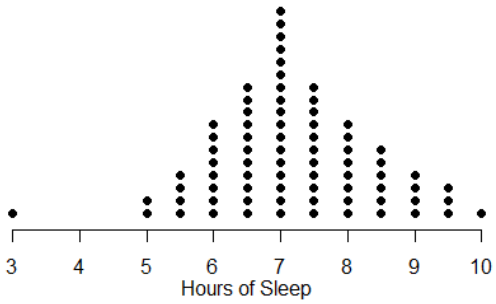
a) Modified version of Item 1 used in the pilot test:
A high school teacher is concerned with whether her students are getting enough sleep. She asked each one of her students to report the number of hours slept the previous night. A plot of the results is given below.



She would like to report her findings to the school principal using appropriate statistical language to describe and interpret the distribution in context. Select the most appropriate statement that she could use to summarize the results.

a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five. The standard deviation is about 2 hours.
b. The distribution of hours of sleep is a normal curve and centered at 7. There is a gap between three and five. The student who slept 3 hours may be an outlier.
c. The distribution of hours of sleep is somewhat bell-shaped. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.

b) Final version of Item 1:
A high school teacher is concerned with whether her students are getting enough sleep. She asked each one of her students to report the number of hours slept the previous night. A plot of the results is given below.



She would like to report her findings to the school principal using appropriate statistical language to describe and interpret the distribution in context. Select the most appropriate statement that she could use to summarize the results.

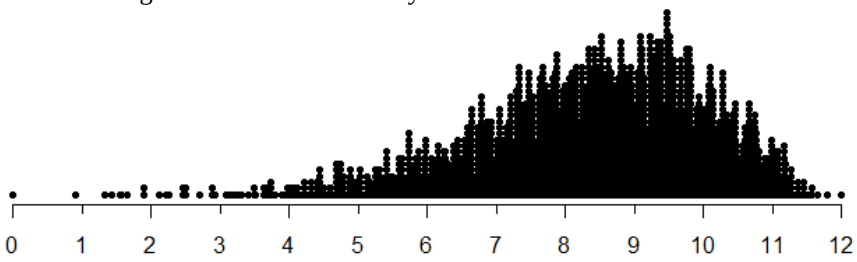
a. The values go from 3 to 10, increasing in height to 7, then decreasing to 10. The most values are at 7. There is a gap between three and five.
b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
c. Many students seem to be getting 7 hours of sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.
d. The distribution of hours of sleep is somewhat normal, with an outlier at 3. The typical amount of sleep is about 7 hours and standard deviation is about 1 hour.

Figure I1. Pilot version and final version of Item 1/1A.

Measures of Center

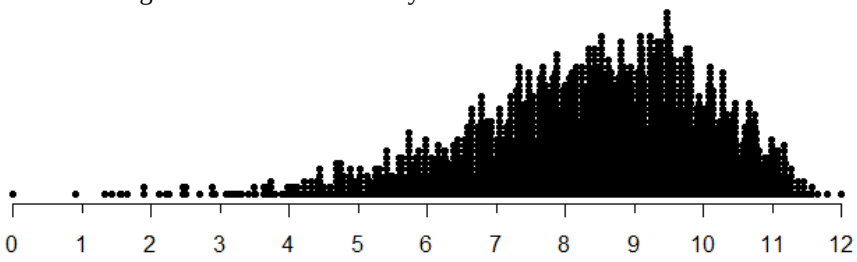
All items in this area had good item characteristics. However, during the pilot test, one student emailed the researcher and expressed difficulties regarding Item 10/1C (Figure I2a). The student reported that the overlap between alternatives b and c made the item inappropriate. This student's feedback was aligned with guideline 22 from Haladyna, Downing, and Rodriguez (2002): "keep choices independent; choices should not be overlapping." Therefore, changes were made to the alternative options. The final version of Item 10/1C is shown in Figure I2b.

a) Version of Item 10 used in the pilot test:
Which of the following intervals is MOST likely to include the mean of the distribution below?



a. 6 to 7
b. 8 to 9
c. 9 to 10

b) Final version of Item 10:
Which of the following intervals is MOST likely to include the mean of the distribution below?



a. 4 to 6
b. 7 to 9
c. 10 to 12

Figure I2. Pilot version and final version of Item 10/1C.

Study Design

This area of learning was composed of eight items: four statistical reasoning and four statistical literacy items. However, the aim for this area of learning was to have three statistical literacy and three statistical reasoning items. Therefore, one statistical reasoning and one statistical literacy item were deleted.

Among all statistical literacy items in this area of learning, Item 18/4D had the lowest item discrimination (0.28) and it was deleted. For the statistical reasoning items, Item 21/8D was not the item with the lowest discrimination; however, this item had presented concerns during the think-aloud interviews with students (students reported that the content addressed by this item was problematic and unclear). Therefore, item 21/8D was excluded.

Hypothesis Testing and p-values.

The hypothesis testing and *p*-value area of learning was comprised of 13 items: seven statistical reasoning and six statistical literacy items. Because the aim for this area of learning was to have five items assessing each of the learning goals, two statistical literacy items and one statistical reasoning item were deleted.

The statistical literacy items that were deleted were the ones with the lowest discriminations: items 29/9F and 23/4F. In addition to having a lower discrimination, Item 29/9F also assessed the same content as Item 32/12F (ability to correctly interpret the *p*-value). Among the statistical reasoning items, Item 27/7F was deleted because it had the lowest discrimination.

Confidence Intervals

This area of learning had the correct number of items needed; however, Item 36/2E (Figure I3a) was the most difficulty item, with only 19% of the students getting this item correct. In addition, this item also presented a very low discrimination (0.04). Table 4.4 shows that for Item 36/2E, alternatives *c* and *d* were the most attractive answers for students to choose with 35% and 40% of the sample choosing these answers, respectively. Both answers *c* and *d* were wrong.

However, they contained the typical confidence interval wording for confidence interval interpretation: "we can say with 95% confidence that …." Therefore, students might be focusing more on how to interpret the confidence interval than to answer the question being asked in the item. With the intent to improve item discrimination and decrease difficulty, alternatives *c* and *d* were modified to no longer have the wording related to interval interpretation. The modified version of Item 36/2E is shown in Figure I3b.

a) Version of Item 36 used in the pilot test:
In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

a. We know that 37% of veterans in the sample have been divorced at least once.
b. We know that 37% of veterans in the population have been divorced at least once.
c. We can say with 95% confidence that 37% of veterans in the sample have been divorced at least once.
d. We can say with 95% confidence that 37% of veterans in the population have been divorced at least once.

b) Final version of Item 36:
In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

a. We can say that 37% of veterans in the sample have been divorced at least once.
b. We can say that 37% of veterans in the population have been divorced at least once.
c. We can say that 95% of veterans in the sample have been divorced at least once.
d. We can say that 95% of veterans in the population have been divorced at least once.

Figure I3a. Pilot version and final version of Item 36/2E.

Probability

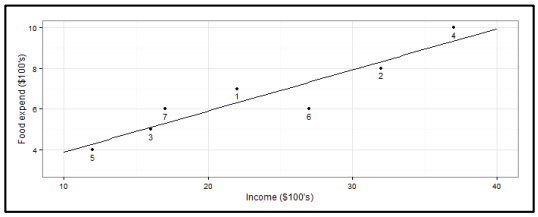
There was a total of six items in this area: three statistical reasoning and three statistical literacy items. The aim for this area of learning was to have four total items (two statistical literacy and two statistical reasoning items). Therefore, there was a need to delete one statistical reasoning and one statistical literacy item.

Item 40/2G (statistical literacy item) and Item 44/6G (statistical reasoning item) were the items with the lowest item discrimination for this area of learning. Therefore, both items were deleted.

Bivariate Data

This area of learning had the correct number of items needed (two statistical literacy and two statistical reasoning items). However, Item 48/5H (Figure I4a) presented low item discrimination (0.14) and it was the fourth hardest item of the instrument (difficulty of 29%). To reduce the complexity of this item, the alternatives were modified to not present numbers but to access what would happen conceptually with intercept and slope. The final version of Item 48/5H can be seen in Figure I4b.

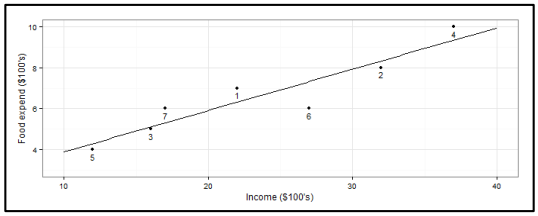
a) Version of Item 48 used in the pilot test:
A random sample of 7 households was obtained, and information on their income and food expenditures for this year was collected. Below is a scatterplot of these data with the regression line superimposed:



The regression equation for these data is: *predicted expend = 1.869 + 0.202 (income)*
After further analysis, the researcher found out that Point 4 was data from another year and decided to exclude it. What would be the regression equation after excluding Point 4?

a. predicted expend = 1.869 + 0.234 (income)
b. predicted expend = 1.869 + 0.202 (income)
c. predicted expend = 2.550 + 0.164 (income)
d. predicted expend = 1.821 + 0.234 (income)

b) Final version of Item 48:
A random sample of 7 households was obtained, and information on their income and food expenditures for this year was collected. Below is a scatterplot of these data with the regression line superimposed:



The regression equation for these data has an intercept of 1.869 and a slope of 0.202:
Predicted expend = 1.869 + 0.202 (income).
After further analysis, the researcher found out that Point 4 was data from another year and decided to exclude it. What would be true about the regression equation after excluding Point 4?

a. Both the intercept and slope will increase.
b. Both the intercept and slope will decrease.
c. The intercept will decrease and the slope will increase.
d. The intercept will increase and the slope will decrease.
e. Both intercept and slope will remain the same.

Figure I4. Pilot version and final version of Item 48.