

RECONCEPTUALIZING STATISTICAL LITERACY: DEVELOPING AN
ASSESSMENT FOR THE MODERN INTRODUCTORY STATISTICS COURSE

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Laura Ann Ziegler

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

Joan Garfield, Adviser
Michelle Everson, Co-adviser

June 2014

© Laura Ann Ziegler, 2014

Acknowledgements

I owe more than I can repay to the many individuals who helped me in the development and completion of this dissertation. First and foremost, I would like to thank my advisor, Dr. Joan Garfield. Her wisdom, innovative thinking and generosity have helped me grow as a researcher and writer. Her guidance and feedback have been invaluable in completing my dissertation work. I am also grateful to have an amazing co-advisor, Dr. Michelle Everson. Her experience and skills as an editor were vital in the writing of my dissertation. Also, I thank Dr. Bob delMas for providing direction in choosing the appropriate analyses, helping me through issues in analyzing my data, and offering instrumental feedback.

I would also like to thank the statistics education community for welcoming me with open arms. I am grateful that multiple statistics educators were willing to serve as reviewers of the statistical literacy assessment I developed for my dissertation and very appreciative that so many statistics instructors were willing to administer my assessment to their students. Without their help, my results would not have been as conclusive.

I am also thankful to the wonderful students and instructors in the statistics education program at the University of Minnesota. Not only did they provide feedback on my writing, but also listened and helped me through difficult times. In particular, I owe my sanity to my duplicate self, Laura Le. I also appreciate the wonderful example of Dr. Andy Zieffler as a researcher and an educator.

To my family, I owe everything. My parents have given me life, encouragement, understanding, kindness, and so much more. I would not be where I am today without

you both. Thank you to my sister and brother for giving me courage by example. I also greatly appreciate the love and support from my in-laws. Last but not least, to my husband and my son Sean, thank you for your patience, forbearance and encouragement. Our times together were welcome and refreshing breaks that brought me so much joy!

Abstract

The purpose of this study was to develop the Basic Literacy In Statistics (BLIS) assessment for students in an introductory statistics course, at the postsecondary level, that includes, to some extent, simulation-based methods. The definition of statistical literacy used in the development of the assessment was the ability to read, understand, and communicate statistical information. Evidence of reliability, validity, and value were collected during the development of the assessment using a mixed-methods approach.

There is a need for a new assessment for introductory statistics courses. Multiple instruments were available to assess students in introductory statistics courses (e.g., Comprehensive Assessment of Outcomes in a First Statistics Course, CAOS; delMas, Garfield, Ooms, & Chance, 2007; Goals and Outcomes Associated with Learning Statistics, GOALS; Garfield, delMas, & Zieffler, 2012); however, there were not assessments available that focused on statistical literacy. In addition, there are introductory statistics courses that are teaching new content such as simulation-based methods (e.g., Garfield et al., 2012; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). To meet the need for a new assessment, the BLIS assessment was developed.

Throughout the development of the BLIS assessment, evidence of reliability, validity, and value were collected. A test blueprint was created based on a review of textbooks that incorporate simulation-based methods (e.g., Catalysts for Change, 2013), reviewed by six experts in statistics education, and modified to provide evidence of validity. A preliminary version of the assessment included 19 items chosen from existing

instruments and 18 new items. To collect evidence of reliability and validity, the assessment was reviewed by the six experts and revised. Additional rounds of revisions were made based on cognitive interviews ($N=6$), a pilot test ($N=76$), and a field test ($N=940$), all of which were conducted with students who had recently completed or were currently enrolled in an introductory statistics course, at the secondary level. Instructors who administered the assessment to their students in the field test completed a survey to gather evidence of the value of the BLIS assessment to statistics educators ($N=26$).

Data from the field test was examined using analyses based on Classical Test Theory (CTT) and Item Response Theory (IRT). When examining individual item scores, coefficient alpha was high, .83. The BLIS assessment contains testlets, so the Partial Credit (PC) model was fit to the data. Evidence of reliability and validity was high; however, more items with high difficulty levels could increase the precision in estimating ability estimates for higher achieving students. Instructors who completed the survey indicated that the BLIS assessment has high value to statistics educators. Therefore, the BLIS assessment could provide valuable information to researchers conducting studies about students' understanding of statistical literacy in an introductory statistics course that includes simulation-based methods.

Table of Contents

Acknowledgements	i
Abstract	iii
List of Tables	xii
List of Figures	xv
Chapter 1 Introduction	1
Rationale for the Study	1
The Basic Literacy In Statistics (BLIS) Assessment	4
Structure of the Dissertation	5
Chapter 2 Review of the Literature	7
Describing Important Learning Outcomes: Statistical Literacy, Reasoning, and Thinking	7
Definitions of Statistical Literacy, Reasoning, and Thinking	8
Comparison of Statistical Literacy, Reasoning, and Thinking	14
Working Definitions of Statistical Literacy, Reasoning, and Thinking	17
Assessing Statistical Literacy	18
Instruments Assessing Statistical Literacy	18
Research on Students' Understanding of Statistical Literacy	27
Statistical Literacy at the Elementary and Secondary School Levels	27
Statistical Literacy at the Postsecondary and Adult Level	31

Summary of Studies of Students' Understanding of Statistical Literacy	35
Teaching Statistical Literacy	36
Courses Designed to Teach Statistical Literacy	37
Summary of Courses Designed to Teach Statistical Literacy	39
Changes in Introductory Statistics: Simulation-based Methods	39
Description of Simulation-based Methods	40
Rationale for Teaching Simulation-based Methods in an Introductory Statistics Course	41
Introductory Statistics Courses and Textbooks that Teach Simulation-based Methods	43
Evaluating Students' Understanding of Statistics in a Simulation-based Introductory Statistics Course	45
Summary of Simulation-based Introductory Statistics Courses	49
Discussion of the Literature Reviewed	51
Summary and Critique of the Literature Reviewed	52
Formulation of the Problem Statement	57
Chapter 3 Methods	59
Determining the Quality of an Assessment	59
Reliability	60
Validity	60
Overview of the Study	62

Test Blueprint Development	65
Development of the Preliminary Test Blueprint	65
Expert Review of the Preliminary Test Blueprint	66
Development of the First Version of the Test Blueprint	68
Assessment Development	69
Item Writing Considerations and Item Characteristics	69
Development of the Preliminary Assessment	71
Expert Review of the Preliminary Assessment	76
Development of the First Version of the Assessment	77
Cognitive Interviews with Students	77
Development of the Second Version of the Assessment	80
Pilot Administration	80
Development of the Third Version of the Assessment	83
Field Test Administration	84
Analysis of Field Test Data	88
Chapter Summary	93
Chapter 4 Results	94
Results from Expert Review of the Test Blueprint	94
Results from Expert Review of the Assessment	98
Results from Student Cognitive Interviews	101
Results from Pilot Test	107
Results from the Field Test	119

Descriptive Statistics	119
Reliability	122
Validity	130
Value	137
Chapter 5 Discussion	139
Summary of the Study	139
Synthesis of the Results	140
Reliability	140
Validity	142
Value	144
Reconceptualizing Statistical Literacy	145
Limitations	146
Implications for Teaching	147
Implications for Future Research	148
Conclusion	151
References	153
Appendix A Version of the BLIS Test Blueprint	176
Appendix B Correspondence with Expert Reviewers of the Preliminary Test Blueprint	189
B1 Invitation to be an Expert Reviewer	189
B2 Examples of Statistical Literacy	191
B3 Preliminary Test Blueprint Review Form	193

Appendix C Versions of the BLIS Assessment	198
C1 Preliminary BLIS Assessment	198
C2 BLIS-1 Assessment	215
C3 BLIS-2 Assessment	230
C4 BLIS-3 Assessment	244
Appendix D Correspondence with Expert Reviewers of the Preliminary Assessment	263
D1 Invitation to Expert Reviewers of the Preliminary Assessment	263
D2 Summary of Changes to the Test Blueprint	265
D3 Preliminary Assessment Review Form	266
Appendix E Correspondence with Instructors and Students for the Cognitive Interviews	293
E1 In-class Invitation Script	293
E2 In-class Sign-up Sheet	294
E3 Email Invitation to Students for the In-person Cognitive Interviews	295
E4 Email Invitation to Students for the Skype Cognitive Interviews	296
E5 Letter Sent to Instructors Requesting Them to Email the Interview Invitation to Their Students	297
E6 Cognitive Interview Protocol	298
E7 Consent Form for Cognitive Interviews	299
Appendix F Correspondence with Instructors and Students for the Pilot Test	302
F1 Invitation to Instructors to Participate in the Pilot Test	302

F2 Instructions for Instructors on How to Administer the BLIS-2 Assessment to Their Students	304
F3 Instructor Survey for the Pilot Test	306
F4 Consent Form for the Pilot Test and Instructions for Students on How to Complete the BLIS-2 Assessment	307
F5 Demographic Questions for Students Taking the BLIS-2 Assessment	309
Appendix G Correspondence with Instructors and Students for the Field Test	310
G1 Invitation Letter to Instructors to Participate in the Field Test Sent Via Personal Emails and the Isolated Statisticians Electronic Mailing List	310
G2 Invitation Letter to Instructors to Participate in the Field Test Sent Via the Consortium for the Advancement of Undergraduate Statistics Education Mailing List	312
G3 Follow-up Letter Sent to Instructors for the Field Test	314
G4 Instructions for Instructors on How to Administer the BLIS-3 Assessment to Their Students	316
G5 Instructor Survey for the Field Test	317
Appendix H Expert Review Results from the Preliminary Test Blueprint	319
Appendix I Expert Review Results from the Preliminary Assessment	340
I1 Changes Made to the Preliminary Assessment	368
Appendix J Student Results from the Cognitive Interviews	383
J1 Changes Made to the BLIS-1 Assessment	396

Appendix K Changes Made to the BLIS-2 Assessment	412
Appendix L Tetrachoric Correlation Residuals from the Single-factor Confirmatory Factor Analysis	431

List of Tables

Table 1 <i>Definitions of Statistical Literacy</i>	15
Table 2 <i>Definitions of Statistical Reasoning</i>	15
Table 3 <i>Definitions of Statistical Thinking</i>	16
Table 4 <i>Percentages of Items Measuring Statistical Literacy</i>	21
Table 5 <i>Courses that Teach Simulation-based Methods</i>	50
Table 6 <i>Overview of Assessment Development, Data Collection, and Analysis</i>	64
Table 7 <i>BLIS Assessment Items Chosen from the CAOS Test, Artist Topic Scale Tests, Artist Item Database, and the GOALS Assessment Matched with Their Learning Outcomes and Sources</i>	73
Table 8 <i>Newly Written BLIS Assessment Items Matched with Their Learning Outcomes and Sources</i>	75
Table 9 <i>Course Characteristics as Reported by Instructors</i>	86
Table 10 <i>Demographic Characteristics of Students Who Took BLIS-3 in the Field Test</i>	88
Table 11 <i>New Item for the BLIS-1 Assessment Created to Replace the Item with Low Ratings from Reviewers for the Preliminary Assessment</i>	100
Table 12 <i>Changes Made to Item 3 After the Student Cognitive Interviews. Words that were Added are Underlined and Words that were Deleted are Crossed Out.</i>	102
Table 13 <i>Items 9 and 17 on the BLIS-1 Assessment that were Changed to be Constructed-Response Items for the BLIS-2 Assessment</i>	104

Table 14 <i>Item 33 on the BLIS-1 Assessment that One Student Struggled with in the Cognitive Interview</i>	107
Table 15 <i>Percentage of Students (N=76) Who Chose Each Selected-Response Option for the 16 Selected-Response Items Administered in the Pilot Test</i>	109
Table 16 <i>Percentage of Students Who Chose Each Selected-Response Option for Items 16 and 27 Conditioned by Course</i>	111
Table 17 <i>Item 35 on the BLIS-2 Assessment</i>	112
Table 18 <i>Item 26 on the BLIS-2 Assessment and Student Explanations for Incorrectly Reporting a P-Value of .04</i>	116
Table 19 <i>Item 29 on the BLIS-2 Assessment Which Did Not Appear to Measure the Intended Learning Outcome</i>	117
Table 20 <i>Percentage of Students (N=940) Who Chose Each Selected-Response Option for All 37 Items Administered in the Field Test</i>	120
Table 21 <i>Total Percentage of students Who Answered Correctly or Incorrectly for Items 29 and 30 on the BLIS-3 Assessment</i>	121
Table 22 <i>Factor Loadings for One-Factor CFA Model with 36 Individual Items</i>	125
Table 23 <i>Fit Indices for One-Factor CFA Models</i>	126
Table 24 <i>Factor Loadings for One-Factor CFA Model with 32 Individual Items and Four Testlets</i>	127
Table 25 <i>Fit Indices for Rasch, 2PL, and PC Models</i>	128
Table 26 <i>Item Parameters for the PC Model with 32 Individual Items and Four Testlets</i>	131

Table 27 <i>Value Ratings Reported by 26 Introductory Statistics Instructors</i>	138
Table A1 <i>Preliminary Test Blueprint</i>	176
Table A2 <i>BLIS Test Blueprint-1</i>	180
Table A3 <i>BLIS Test Blueprint-2</i>	183
Table A4 <i>BLIS Test Blueprint-3</i>	186
Table H1 <i>Ratings for the Learning Outcomes in the Preliminary Test Blueprint from the Six Expert Reviewers. Groups of Similar Ratings are Also Included Where Group A is the Highest Rated Group and Group D is the Lowest Rated Group.</i>	319
Table H2 <i>Comments from Reviewers on Specific Learning Outcomes and Changes Made to the Preliminary Test Blueprint</i>	326
Table I1 <i>Ratings for the Items in the Preliminary Assessment from the Six Expert Reviewers. For Each Item, Reviewers were Asked How Much They Agreed or Disagreed with the Following Statement: “The assessment item measures the specified learning outcome.” Groups of Similar Ratings are Also Included Where Group A is the Highest Rated Group and Group D is the Lowest Rated Group.</i>	340
Table I2 <i>Comments from Reviewers on Specific Items</i>	342
Table J1 <i>Comments from Students that were Used to Make Changes to Assessment Items to Create the BLIS-2 Assessment</i>	383
Table L <i>Tetrachoric Correlation Residuals from the Single-factor Confirmatory Factor Analysis</i>	431

List of Figures

<i>Figure 1.</i> Example item on the SRA test that measures statistical literacy.	21
<i>Figure 2.</i> Example item on the CAOS test that measures statistical literacy.	23
<i>Figure 3.</i> Example item on the ARTIST Data Representation Topic Scale test that measures statistical literacy.	24
<i>Figure 4.</i> Example item on the GOALS test that measures statistical literacy.	25
<i>Figure 5.</i> Example item on the ARTIST Measures of Center Topic Scale test that measures statistical reasoning.	54
<i>Figure 6.</i> Dotplot of students' total scores for the 36 items on the BLIS-3 assessment.	122
<i>Figure 7.</i> Scree plots of eigenvalues for the BLIS-3 assessment.	124
<i>Figure 8.</i> Item information curves for the 32 items and 4 testlets.	129
<i>Figure 9.</i> Test information function and standard error of measurement for the BLIS-3 assessment.	130
<i>Figure 10.</i> Item characteristic curves of 28 items and 4 testlet-based items. For the 28 items, a value of 1 represents an incorrect response and a value of 2 represents a correct response. For the 4 testlet-based items, a value of 1 represents an incorrect response, a value of 2 represents a partially correct response, and a value of 3 represents a correct response.	132

Chapter 1

Introduction

Statistical literacy has been described as an important learning outcome in introductory statistics courses (Garfield, delMas, & Zieffler, 2010). A full, working definition of statistical literacy was needed in order to guide the direction of the literature review and the development of the assessment described in this paper. After examining the literature, it was found that there was little consensus on the definition of statistical literacy. Statistical literacy has been defined as knowing the basic language of statistics (Garfield, delMas, & Chance, 2002), but also as communicating, interpreting, and being critical of statistical information (Gal, 2002). The definition of general literacy, the ability to read and write (“Literacy,” n.d.a; “Literacy,” n.d.b), was used here to provide an argument for defining statistical literacy as being able to read, understand, and communicate statistical information. To elaborate on this definition, some components of statistical literacy include understanding definitions and terms, creating graphs, and interpreting visual representations of data.

Rationale for the Study

There are multiple assessments that measure students understanding of statistics in an introductory statistics course at the postsecondary level: Statistics Reasoning Assessment (SRA; Garfield, 2003), Statistics Concepts Inventory (SCI; Reed-Rhoads, Murphy, & Terry, 2006), Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS; delMas et al., 2007), Assessment Resource Tools for Improving Statistical Thinking (ARTIST; Garfield et al., 2002) Topic Scale tests, Quantitative Reasoning Test-Version 9 (QR-9; Sundre, Thelk, & Wigtil, 2008), and Goals and

Outcomes Associated with Learning Statistics (GOALS; Garfield et al., 2012). Each assessment contains some items measuring statistical literacy, but many of these assessments focus on higher order outcomes, such as being able to make connections and reason about statistics. Each of these assessments, except for the GOALS assessment, was designed for students in an introductory statistics course taught with normal-based methods, or what has been termed the *consensus curriculum* (Cobb, 2007). Further, although the GOALS assessment was designed for students in an introductory statistics course taught with simulation-based methods, it does not contain very many items that measure statistical literacy.

Multiple studies have been conducted to examine students' understanding of statistical literacy (e.g., Pierce & Chick, 2013; Schield, 2006; Watson, 2011). These studies involved elementary children up to adult learners. Multiple statistical learning outcomes were examined in the studies, such as understanding basic terminology, interpreting data presented in tables and graphs, and understanding percentages and probabilities. Results from the studies were mixed; however, a majority of the studies demonstrated that students do not have a very high level of statistical literacy (Anderson, Gigerenzer, Parker, & Schulkin, 2014; Galesic & Garcia-Retamero, 2010; Pierce & Chick, 2013; Schield, 2006; Turegun, 2011; Wade, 2009; Watson, 2011). There were studies that showed evidence that students struggled with being able to make interpretations of statistical results (Jones et al., 2000; Ridgway, Nicholson, McCusker, 2008; Yolcu, 2012). There were two studies that found students could successfully describe data (Jones et al., 2000; Sharma, Doyle, Shandil, & Talakia'atu, 2012), and students have been shown to be able to understand tables of percentages (Atkinson,

Czaja, & Brewster, 2006). Unfortunately, a majority of these studies did not take advantage of well-established assessments, such as the assessments mentioned previously. Only one study used a high-quality assessment (ARTIST Topic Scale) that was created for research purposes (Turegun, 2011). None of the studies used assessments that had evidence of validity or reliability. Overall, these studies provide some insight to students' statistical understanding, but there is there is more work to be done.

The content taught in introductory statistics courses have changed greatly over the years (Chance & Garfield, 2002; Cobb, 1992; Moore, 1997; Utts, 2003). One change of interest is that simulation-based methods, such as randomization tests, are being taught in introductory statistics courses in addition to or in lieu of normal-based methods, such as the *t*-test (e.g., Garfield et al., 2012; Tintle et al., 2011). This change has occurred for many reasons, including the arguments that simulation-based methods are easily grasped (Cobb, 2007) and promote understanding (Hesterberg, Monaghan, Moore, Clipson, & Epstein, 2003). Multiple new introductory statistics textbooks are being published that incorporate simulation-based methods (e.g., Catalysts for Change, 2013; Lock, Lock, Lock Morgan, Lock, & Lock, 2013). As mentioned previously, the only assessment available that was designed for research purposes in an introductory statistics course taught with simulation-based methods is the GOALS assessment, and this includes very few items that measure statistical literacy. Therefore, new assessments of statistical literacy could include topics that are taught in introductory statistics courses that use simulation-based methods as well as those courses that do not use simulation-based methods.

Research has been conducted with students' in simulation-based introductory statistics courses in order to examine what statistical knowledge students gain in their course (e.g., Garfield et al., 2012; Holcomb, Rossman, & Chance, 2011; Tintle, et al., 2011). This research has demonstrated that, when compared to students who complete a normal-based course, students in a simulation-based course have a better understanding of some statistical concepts, especially statistical inference. However, given the dearth in assessments for students in simulation-based introductory statistics courses, a majority of the studies examined in this paper used either instructor-made assessments or assessments designed for students in a normal-based introductory statistics course.

Considering the move to include simulation-based methods in introductory statistics courses, a new assessment of statistical literacy is clearly needed, and research involving students in simulation-based introductory statistics courses could benefit from an assessment of statistical literacy that has evidence of validity and reliability. Such an assessment could be used with students as a pretest and posttest in an introductory statistics course at the postsecondary level to see what statistical literacy knowledge students gained during the course. A new assessment of statistical literacy could also be used to compare methods of teaching statistics to students in a simulation-based introductory statistics course. For example, a comparison could be made between courses that teach simulation-based methods first and then normal-based methods with courses that teach the two methods together throughout the course. Therefore, a new assessment of statistical literacy is needed to examine students' knowledge in studies such as the ones mentioned here.

The Basic Literacy In Statistics (BLIS) Assessment

In this dissertation, the development of a new assessment, The Basic Literacy In Statistics (BLIS) assessment, is described. Based on recommendations from AERA, APA, and NCME (1999), evidence of reliability and validity were collected throughout the development process. Sources of reliability and validity evidence that were collected include expert reviews of the BLIS test blueprint, expert reviews of the assessment, cognitive interviews with students, a small-scale pilot of the assessment, and a large-scale field test of the assessment. Analyses of the field test data were conducted based on Classical Test Theory (CTT) and Item Response Theory (IRT).

Structure of the Dissertation

In chapter 2, a detailed literature review is provided to further explain the need for a new assessment of statistical literacy. Definitions of statistical literacy that have been cited in the literature, as well as definitions of other terms such as statistical reasoning and statistical thinking, are compared to give justification for the definition of statistical literacy used in this study. Existing instruments that include items measuring statistical literacy are then discussed. Studies of students' understanding of statistical literacy are critiqued, and this is followed by a description of courses and textbooks designed to teach statistical literacy. The rationale for including simulation-based methods in introductory statistics courses is described, and the structure of these new courses (and the new textbooks that go along with these courses) is discussed. Studies that have been conducted with students in a simulation-based introductory statistic course are also detailed to provide further justification for a new assessment of statistical literacy.

Chapter 3 includes the methodology used to create the BLIS assessment. The methods consisted of how the assessment was developed as well as how evidence of validity and reliability was collected. Lastly, a description of the analysis that was conducted with the large-scale field test data is described.

Chapter 4 presents the results of the study. This includes the results from the expert reviews of both the test blueprint and assessment. Results from student cognitive interviews and the small-scale pilot are described followed by a presentation of the results from the analysis conducted with the large-scale field test data.

In chapter 5, a discussion is provided, along with the conclusions drawn based on the results of this study. The chapter ends with limitations of the study, and implications for teaching and implications for future research.

Chapter 2

Review of the Literature

The purpose of this critical review of the literature was to try to understand and describe statistical literacy, how it has been taught in introductory college-level statistics courses, and to re-conceptualize what statistical literacy means in the context of new curriculum based on simulation methods. In addition, this review showed the need for a new assessment of statistical literacy for students in a simulation-based introductory statistics course. First, definitions of statistical literacy, reasoning, and thinking were compared in order to create a working definition of statistical literacy to guide the review of the literature. Using this definition as a foundation, the following were summarized and critiqued: assessments of statistical literacy, studies about students' understanding of statistical literacy, and courses that teach statistical literacy. In light of current changes in the content taught in introductory statistics courses, the case was made for new assessments of statistical literacy.

Describing Important Learning Outcomes: Statistical Literacy, Reasoning, and Thinking

Different terms have been used to describe various student outcomes in an introductory statistics course. Three outcomes related to students' understanding of statistics are statistical literacy, statistical reasoning, and statistical thinking. The first effort to describe and distinguish between these outcomes was made by Garfield et al. (2002) in order to create assessment items aligned with each of these outcomes. There is not a consensus among statistics educators regarding the definitions of these outcomes (Ben-Zvi & Garfield, 2004). In this section, a selection of commonly cited definitions of

statistical literacy, reasoning, and thinking were described in order to come up with a working definition of statistical literacy to guide the review and discussion of the literature.

Definitions of statistical literacy, reasoning, and thinking.

General literacy. In order to understand the nature of statistical literacy, it was helpful to look at definitions of literacy in general. Two dictionary definitions of literacy are the ability to read and write (“Literacy,” n.d.b) and having basic skills of a subject, such as computer literacy (“Literacy,” n.d.a). According to the Ohio Literacy Resource Center (2012), these definitions were too simplistic for today’s world. For example, the Educational Development Center (n.d.) stated that literacy is when a person is able to use reading and writing in “shaping the course of his or her own life” (para. 1). As these definitions become more developed, they can include other skills.

There are definitions of literacy that include mathematical and statistical capabilities. The National Literacy Act of 1991 established the National Institute for Literacy and defined literacy as “an individual’s ability to read, write, and speak in English, and compute and solve problems at the levels of proficiency necessary to function on the job and in society, to achieve one’s goals, and develop one’s knowledge and potential” (H. R. 751-102nd Congress, 1991, p. 7). Knowing how to use technology has also been included in definitions of literacy (e.g., Ohio Literacy Resource Center, 2012). Another definition of literacy includes the “ability to identify, understand, interpret, create, communicate and compute using printed and written materials associated with varying contexts” (The United Nations Educational, Scientific and Cultural Organization, 2003). According to the World Literacy Foundation, literacy also

includes critical thinking (n.d., para. 1). The Organisation for Economic Co-operation and Development described literacy as understanding written texts that include visuals such as graphs (OECD, 2012). Each definition of general literacy provided here includes the basic ability to read and write but varies in the additional skills that were included such as thinking critically. It is not surprising that there is not a consensus on the definition of statistical literacy because there are so many definitions of general literacy.

Statistical literacy. Just as definitions of general literacy ranged from being able to read and write to being able to think critically, definitions of statistical literacy also differed in the spectrum of basic skills to critical thinking. One of the first published definitions of statistical literacy was made by Walker (1951), who served as president of the American Statistical Association in 1944 and the American Educational Research Association from 1949 to 1950 (American Statistical Association, n.d.). In order to define statistical literacy, Walker examined definitions of general literacy that included the ability to read and write, and she suggested that statistical literacy is the ability to communicate statistical information. Chick and Pierce (2013) claimed that statistical literacy encompasses general literacy, numeracy, statistics, and data presentation which includes the ability to reason with information presented in graphs and tables. Several statistics educators have described statistical literacy as understanding and using the basic language of statistics (Garfield et al., 2005; Garfield & delMas, 2010; Garfield et al., 2002; Lehohla, 2002).

In contrast, statistical literacy has been defined by others as including higher order skills such as the ability to interpret and critically evaluate reported statistics (Jordan, 1981; Sanchez, 2010). These definitions have been combined to define statistical literacy

as communicating, interpreting, and being critical of statistical information (Gal, 2002, 2011; Schield, 1999; Smith, 2002). Statistics educators, such as Martinez-Dawson (2010) and Schield (1999), emphasized the role of the statistical information people come across in the real world. Wallman (1993) also incorporated the ability to understand and critically evaluate statistical information in the real world but added that a statistically literate citizen should be able to appreciate contributions that statistical thinking provides to make decisions.

Instead of providing a specific definition of statistical literacy, there are statistics educators who chose to describe what statistical literacy is by providing a list of ideas and skills. In the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report for pre-K-12, statistical literacy was described as being able to understand polls, understand the behavior of random samples, interpret a margin of error, make daily personal choices, and understand and question scientific findings (Franklin et al., 2007). Kaplan and Thorpe (2010) assembled a similar list of skills for college students. They studied five publications (Cobb, 1992; Franklin et al., 2007; Gal, 2002; Rumsey, 2002; Utts, 2003) to come up with five topics statistically literate students should understand: data and experimental design, probability, variability, descriptive statistics, and conclusions and inference.

Some scholars have tried to describe statistical literacy using different categories. One of the major contributors to the literature on statistical literacy is Jane Watson. She suggested that there are three tiers of statistical literacy: understanding of basic statistical problems and terminology, being able to use the basics in the real world, and questioning statistical conclusions and results (Watson, 2011). Six constructs of statistical literacy

were defined by Watson and Callingham (2003): critical/mathematical, critical, consistent/non-critical, inconsistent, informal, and idiosyncratic. Analysis based on item response theory was used by Callingham and Watson (2005) to provide evidence of validity for the six constructs. Callingham and Watson administered a statistical literacy test to 673 students in Grades 5 to 10 and found that there were gaps in the difficulty levels of the items which allowed the items to be categorized into the six constructs. In contrast, Rumsey (2002) reviewed multiple definitions of statistical literacy and claimed that the term was too broad. As a result, she split statistical literacy into two separate categories: “statistical competence” which corresponded to the basic knowledge of statistical reasoning and thinking and “statistical citizenship” which was the statistical ability to function in today’s society.

Statistical reasoning. Statistical reasoning can be defined similarly to general reasoning. According to the Merriam-Webster dictionary, reasoning is defined as “the use of reason; especially: the drawing of inferences or conclusions through the use of reason” (“Reasoning,” n.d.). There are statistics educators who have defined statistical reasoning as a step beyond the basics; students are able to reason with statistical ideas and understand statistical information (Garfield & delMas, 2010; Garfield et al., 2002; Garfield & Gal, 1999). It includes making sense of statistical information, constructing interpretations, and making connections between ideas and topics in statistics. Statistical reasoning has also been defined as what a student is able to do with statistical content, and some argued that it includes three stages: comprehension, planning and execution, and evaluation and interpretation (Chervany, Collier, Fienberg, Johnson, & Neter, 1997). Instead of looking at stages, Jones, Langrall, Mooney, and Thornton (2004) listed four

processes: describing data, organizing and reducing data, representing data, and analyzing and interpreting data. Lastly, Lovett (2001) described statistical reasoning as being able to understand and conduct statistical analyses and being able to summarize, draw conclusions, and make predictions.

Statistical thinking. According to Wild and Pfannkuch (1999), statistical thinking is an all-encompassing understanding of statistics. In another description, Cobb (1998) described two continuums that are a part of statistical thinking: computational/algorithmic thinking to logical/deductive thinking and verbal/interpretation to graphical/dynamic. Cobb and Moore (1997) claimed that statistical thinking includes context, data production, data analysis, and formal inference. Another perspective provided by De Veaux and Velleman (2008) was that statistical thinking involves seven unnatural acts: think critically, be skeptical, think about variation, focus on what we don't know, perfect the process, think about conditional probabilities and rare events, and embrace vague concepts.

Statistics educators who categorized statistical knowledge into the three categories of statistical literacy, reasoning, and thinking, claimed that statistical thinking is a higher order of thinking than the other two (Garfield & delMas, 2010; Garfield et al., 2002). Statistical thinking has been described as understanding how and why, knowing when to inspect and explain variability, understanding the data, and being able to associate the data with the appropriate analysis to investigate a problem. It includes being able to critically evaluate statistical reports. In general, statistical thinking means being able to think like a statistician.

There are statistics educators who have used the terms statistical literacy, reasoning, and thinking interchangeably. As mentioned previously, Watson (2011) defined statistical literacy by describing three tiers. In Watson's 1997 article, she used those same three tiers to describe statistical thinking. In addition, statistical thinking was defined as describing data, organizing and reducing data, representing data, and analyzing and interpreting data by Jones et al. (2000) and Mooney (2002). As mentioned earlier, these same four processes were described as statistical reasoning by Jones et al. (2004).

Related terms. It should be noted that there are other terms that include aspects of statistical understanding such as quantitative literacy, quantitative reasoning, quantitative practices, and numeracy. While definitions of these terms often include some statistics topics, they usually focus primarily on mathematical learning outcomes.

Quantitative literacy includes the basics of statistics, reasoning, logic, and evaluating risks (Kolata, 1997). People who are quantitatively literate are able to “reason in numerical, data, spatial, and chance settings” (Dossey, 1997, p. 48). The National Adult Literacy Survey (NALS) was created to measure literacy, and in this survey, literacy was broken into three scales: prose literacy, document literacy, and quantitative literacy (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993). Kirsch et al. claimed that quantitative literacy includes skills such as being able to balance a checkbook and calculate the amount of interest for a loan.

There are people who chose to use other terminology rather than *literacy*. For example, when asked to define quantitative literacy, Cobb (1997) and Denning (1997) claimed that we should not use the word *literacy* because it implies we are talking about reading and writing. Instead, Cobb (1997) described *quantitative reasoning* on a

continuum; the low end is calculating, the middle is solving a problem in an applied context, and at the high end is reasoning about relationships. Denning (1997) described *quantitative practices* as working with numbers, uncertainty, designing experiments, modeling, making conclusions, and more.

The term *numeracy* goes back to the Crowther Report (Crowther, 1959) when it was described as understanding the scientific approach which includes hypothesis and experiments as well as being able to think quantitatively. In addition, numeracy has been claimed to be more than mathematics; it includes many life skills such as comprehending polls (Gal, 1997). Numeracy has also been described as possessing mathematical skills as well as modeling, interpreting, evaluating/analyzing, communicating, and understanding relationships, data, and chance (OECD, 2012). Chick and Pierce (2013) give a different perspective by saying that numeracy is one of the components included in statistical literacy.

While quantitative literacy, quantitative reasoning, quantitative practices, and numeracy include both mathematical and statistical abilities, the focus of this paper is on statistical literacy. Therefore, the remainder of this discussion will include only statistical literacy, statistical reasoning, and statistical thinking.

Comparison of statistical literacy, reasoning, and thinking. When comparing the definitions for statistical literacy, reasoning, and thinking, it appears that there is a lot of overlap. Tables 1, 2, and 3 include a comparison of definitions for each of these three outcomes. The terms statistical literacy and statistical thinking have sometimes been used interchangeably and have both been defined as including the big ideas, such as being able to critically evaluate statistical reports.

Table 1

Definitions of Statistical Literacy

Publication	Understand and comprehend	Communicate and interpret	Make inferences	Plan and execute procedures	Use logic and reason	Apply to real life contexts	Question and criticize	Consider how and why	Appreciate statistical contributions
Franklin et al. (2007), Martinez-Dawson (2010), Schield (1999), Watson (2011)	X	X	X	X	X	X	X	X	
Gal (2002, 2011), Jordan (1981), Sanchez (2010), Smith (2002)	X	X	X		X	X	X		
Wallman (1993)	X				X	X	X		X
Chick & Pierce (2013)	X	X	X		X	X			
Walker (1951)	X	X				X			
Garfield et al. (2005), Garfield et al. (2002)	X	X				X			
Lehohla (2002)	X					X			

Table 2

Definitions of Statistical Reasoning

Publication	Understand and comprehend	Communicate and interpret	Make inferences	Plan and execute procedures	Use logic and reason	Apply to real life contexts
Chervany et al. (1997), Jones et al. (2004), Lovett (2001)	X	X	X	X	X	X
Garfield et al. (2002), Garfield & Gal (1999)		X	X		X	X

Table 3

Definitions of Statistical Thinking

Publication	Understand and comprehend	Communicate and interpret	Make inferences	Plan and execute procedures	Use logic and reason	Apply to real life contexts	Question and criticize	Consider how and why	Think like a Statistician	Appreciate statistical contributions
Wild and Pfannkuch (1999)	X	X	X	X	X	X	X	X	X	X
Cobb (1998)	X	X	X	X	X	X	X	X	X	
Watson (1997)	X	X	X	X	X	X	X	X		
Jones et al. (2000), Mooney (2002)	X	X	X	X	X	X				
Garfield et al. (2002)						X	X	X	X	

The three outcomes have also been described as forming a hierarchy going from statistical literacy to reasoning to thinking (Garfield & Ben-Zvi, 2007; Garfield & delMas, 2010; Garfield, delMas, & Chance, 2003). Statistical literacy was often described as understanding the basic language of statistics, reasoning as a step beyond that where people can reason with statistics, and thinking as the highest ability. There appears to be overlap between the definitions of statistical literacy and statistical reasoning provided by Garfield and her colleagues. For example, definitions for both statistical literacy and statistical reasoning included being able to make interpretations. Statistical literacy was described as being able to “interpret representations of data” and statistical reasoning was described as being able to “fully interpret statistical results” (Garfield & delMas, 2010, p. 3). Statistical results could include results from inferential analyses.

Watson’s (1997, 2011) three tiers of statistical literacy are similar to the three definitions presented by Garfield and her colleagues. As mentioned previously, Watson described statistical literacy (2011) and statistical thinking (1997) using the same three tiers. When asked in an interview what she felt the difference was between statistical literacy and thinking, she proposed that statistical thinking is a little bit broader than statistical literacy. Statistical literacy is specifically about critical thinking and statistical thinking is more general (personal communication, November 29, 2012). Clearly there is no consensus on definitions of statistical literacy, reasoning, and thinking.

Working definitions of statistical literacy, reasoning, and thinking.

Considering the various definitions found in the literature, the definitions of statistical literacy, reasoning, and thinking that were chosen to build off of were the definitions described as three levels of cognitive outcomes (Garfield & Ben-Zvi, 2007; Garfield &

delMas, 2010; Garfield et al., 2003; Garfield & Franklin, 2011). For the purpose of this paper, a full, working definition of statistical literacy was needed. Drawing on the basic definition of literacy, which is the ability to read and write (“Literacy,” n.d.a; “Literacy,” n.d.b), statistical literacy includes the ability to read, understand, and communicate statistical information. Statistical reasoning is about making connections and going beyond basic statistical literacy. The highest level would be statistical thinking, which is being able to think like a statistician. Statistical literacy is the only term that has been used to refer to a basic understanding of statistics, and because more studies are needed that look at a basic understanding of statistics, the definition used in the present study is the ability to read, understand, and communicate statistical information. The focus in the next section is on statistical literacy and how to assess it based on this working definition of statistical literacy.

Assessing Statistical Literacy

Assessments of students’ statistical literacy can take on many forms, such as exams in class, student presentations, written assignments, or even large-scale assessments. Large-scale assessments are often used for research purposes and will be the focus of this section.

Instruments assessing statistical literacy. There currently exist multiple assessment instruments of learning outcomes for introductory statistics courses. Ten standard and widely used assessments were selected to be described in this paper. These include: National Assessment of Educational Progress (NAEP), Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), Advanced Placement (AP) Statistics Exam, Statistics Reasoning Assessment

(SRA), Statistics Concepts Inventory (SCI), Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS), Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Topic Scale tests, Quantitative Reasoning Test-Version 9 (QR-9), and Goals and Outcomes Associated with Learning Statistics (GOALS).

Instruments for elementary and secondary students. The *National Assessment of Educational Progress* (NAEP; National Center for Education Statistics, 2011a), *Program for International Student Assessment* (PISA; National Center for Education Statistics, 2011b), and *Trends in International Mathematics and Science Study* (TIMSS; National Center for Education Statistics, 2011c) all measure mathematical understanding as well as statistical understanding. Each consists of a combination of selected-response and constructed-response items with the exception of the TIMSS version for fourth graders, which is selected-response only. The TIMSS is administered every four years to fourth and eighth grade students internationally. There are about 37 mathematics items on the fourth grade exam and 79 on the eighth grade exam, with approximately 15% of the items being statistics items. The NAEP is administered to twelfth grade students in addition to fourth and eighth grade students in the United States every year. The fourth grade mathematics portion consists of approximately 10% statistics items, the eighth grade about 15% statistics items, and the twelfth grade around 25% statistics items. The NAEP has been used to measure student growth in understanding statistics (Shaughnessy, 2007). The PISA mathematics assessment is administered every nine years to fourth and eighth graders internationally and attempts to measure literacy using real-life situations (McGrath, 2008). As a result, out of approximately 85 items, around 40% are statistics items. The PISA mathematics assessment consists of four subscales: space and shape,

change and relationship, quantity, and uncertainty. The items in the uncertainty subscale and a subset of items in the change and relationship subscale measure statistical literacy (François, Monteiro, & Vanhoof, 2008).

The *AP Statistics Exam* is an assessment instrument that is specific for secondary school students (CollegeBoard, 2010). The AP Statistics Exam is used to assess whether students have obtained enough statistical knowledge to gain credit for an introductory statistics course at a college of their choice. The AP Statistics Exam consists of 40 selected-response items as well as six free-response items that are meant to measure statistical literacy and reasoning. The AP Statistics Exam items are changed year to year.

Instruments for postsecondary students. There are six instruments presented in this paper that have been developed and used with students at the postsecondary level: Statistics Reasoning Assessment (SRA), Statistics Concept Inventory (SCI), Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS), Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Topic Scale tests, Quantitative Reasoning Test-Version 9 (QR-9), and Goals and Outcomes Associated with Learning Statistics (GOALS). The assessments described in this subsection are presented in the order they were created.

The *Statistics Reasoning Assessment* (SRA) can be used to determine whether or not students are reasoning about statistics and probability (Garfield, 2003). The SRA consists of 20 selected-response items. It is different from many other assessments in that it is not just looking for correct answers. Each item consists of multiple correct answers, but only one includes the correct thinking or rationale. This assessment has been translated into other languages and used internationally. Using the working definition of

statistical literacy presented in this paper and review of the literature, it was determined that one fifth of the items measure statistical literacy (see Table 4). See Figure 1 for an example item that measures statistical literacy.

Table 4

Percentages of Items Measuring Statistical Literacy

Assessment	Number of Items	Percent
SRA	20	20.0
SCI	25	60.0
CAOS	40	35.0
ARTIST Topic Scales ^a	118	64.4
GOALS	27	11.1

^aThe eleven ARTIST Topic Scales were combined to create this summary.

8. Two containers, labeled A and B, are filled with red and blue marbles in the following quantities:

Container	Red	Blue
A	6	4
B	60	40

Each container is shaken vigorously. After choosing one of the containers, you will reach in and, without looking, draw out a marble. If the marble is blue, you win \$50. Which container gives you the best chance of drawing a blue marble?

___ a. Container A (with 6 red and 4 blue)
 ___ b. Container B (with 60 red and 40 blue)
 ___ c. Equal chances from each container

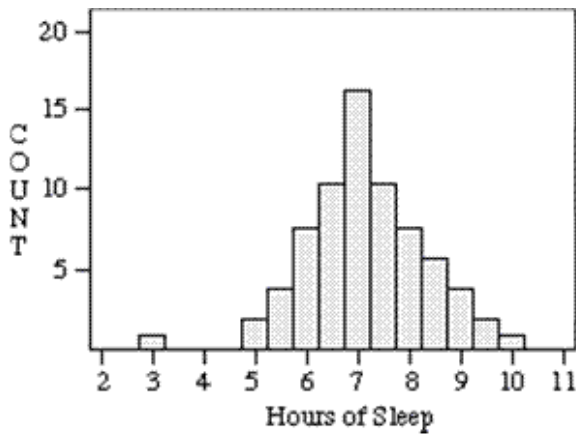
Figure 1. Example item on the SRA test that measures statistical literacy.

The *Statistics Concepts Inventory* (SCI) measures conceptual understanding of statistics (Reed-Rhoads et al., 2006). It is a selected-response statistics assessment that is often used for engineering and mathematics students. Students in other areas have also taken the assessment but have been shown to score significantly lower on the assessment.

It is unknown if this is because of the students' background or because of characteristics of the assessment itself. The assessment contains four sections: descriptive, probability, inferential, and graphical. After revisions by Allen (2006), there remained a total of 25 items on this assessment. The section with the most items is the descriptive section with nine items; the section with the least amount of items is the graphical section with four items. Based on an inspection of the items, nine of the items have no real-world context and a majority of the items measure statistical literacy (see Table 4). The working definition of statistical literacy and the previous review of the literature were referred to in order to determine which items measure statistical literacy.

The *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS) test was created to look at students' overall statistical knowledge (delMas et al., 2007). The intent was to measure students' conceptual understanding of the big ideas in statistics, including understanding variability. Some of the items on the CAOS test were based on items from the ARTIST Topic Scale tests. The assessment consists of 40 selected-response items which were created over a span of three years and have been tested and revised multiple times. The reliability was relatively high; coefficient alpha was .82. As of April, 2014, over 36,000 students in a college-level introductory statistics course, including students in an AP Statistics course, have taken the CAOS test as a posttest. The test was created to assess statistical literacy and reasoning; however, fewer than half of the items were judged to measure statistical literacy (see Table 4). See Figure 2 for an example item that measures statistical literacy.

The following graph shows a distribution of hours slept last night by a group of college students.



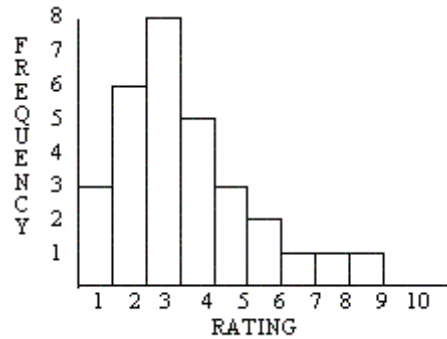
1. Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.
 - a. The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is 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. Most students seem to be getting enough 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 symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours.

Figure 2. Example item on the CAOS test that measures statistical literacy.

The *ARTIST Topic Scales* tests are 11 selected-response tests created to assess the big ideas in statistics such as Data Collection, Probability, and Tests of Significance (Garfield et al., 2002). Each Topic Scale test consists of nine to fifteen selected-response items. The items were created to assess statistical literacy and reasoning (delMas, Garfield, & Ooms, 2005). When looking at the items and using the working definition of statistical literacy, a majority of items appeared to measure statistical literacy (see Table 4). See Figure 3 for an example item from the Data Representation Topic Scale test.

One of the items on the student survey for an introductory statistics course was "Rate your aptitude to succeed in this class on a scale of 1 to 10" where 1 = Lowest Aptitude and 10 = Highest Aptitude. The instructor examined the data for men and women separately. Below is the distribution of this variable for the 30 women in the class.

12. How should the instructor interpret the women's perceptions regarding their success in the class?



- A majority of women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
- The women in the class see themselves as having lower aptitude for statistics than the men in the class.
- If you remove the three women with the highest ratings, then the result will show an approximately normal distribution.

Figure 3. Example item on the ARTIST Data Representation Topic Scale test that measures statistical literacy.

The *Quantitative Reasoning Test-Version 9* (QR-9) is an assessment that is not specific to a course and includes items measuring statistical literacy and reasoning (Sundre et al., 2008). The test can be used to inform mathematics, science, and statistics instructors. The QR-9 has two scales. In the first scale, students “use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomena” and in the second scale, students “discriminate between association and causation, and identify the types of evidence used to establish causation” (p. 5). There are a total of 26 items. Sixteen items measure the first scale, five measure the second scale, and five measure both scales. The assessment is not freely available to the public so it is not known how

many of the items measure statistical literacy. Based on the descriptions of the scales, it seems likely that there could be statistical literacy items in the first scale.

The *Goals and Outcomes Associated with Learning Statistics (GOALS)* test assesses statistical literacy and reasoning in a simulation-based course. The assessment contains 27 selected-response items which are in the process of being revised. Eighteen of the items were based on CAOS items and the other nine items were created to focus on inference using simulation methods. Based on an examination of the items and using the working definition of statistical literacy, there were only three items that measured statistical literacy (see Table 4). See Figure 4 for an example item. An earlier version of the GOALS test was described by Garfield et al. (2012).

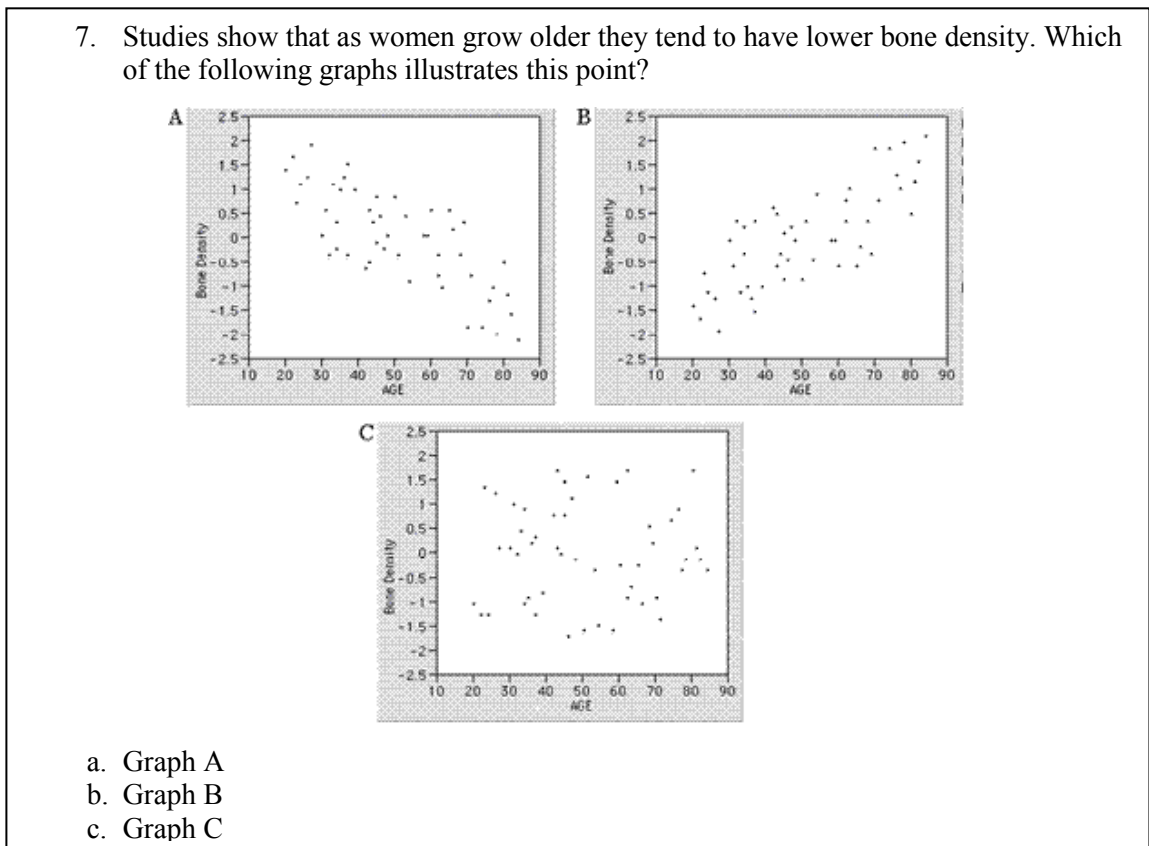


Figure 4. Example item on the GOALS test that measures statistical literacy.

Summary of assessments of statistical literacy. The previously described assessments of learning outcomes for introductory statistics each have various purposes, advantages, and drawbacks. The NAEP, PISA, TIMSS, and AP Statistics exams are for elementary and secondary school students and are not publicly available. As mentioned, the SRA is unique in that it is meant to measure reasoning and not just correctness. The QR-9 is not specific to a statistics course and is not freely available. There are assessments that do not have many items that measure statistical literacy such as the GOALS assessment, SRA, and possibly the QR-9. A real-world context is important to include in measuring statistical literacy (Watson, 2011) and the SCI has nine items out of 25 that have no real-world context. The ARTIST Topic Scales tests and the SCI have more items measuring statistical literacy than the CAOS test, but the CAOS test has more validity evidence and a large national sample that can be used for comparison purposes in teaching experiments. Therefore, if one is planning to assess statistical literacy for college students in introductory statistics courses, the SCI, ARTIST Topic Scales tests, and CAOS assessments are the best options that are widely available. Unfortunately, however, these assessments are becoming outdated.

Statistics educators are now changing what is being taught in introductory statistics courses. Non-parametric procedures such as randomization tests are being taught in addition to or in lieu of normal-based tests such as the t -test (Garfield et al., 2012; Tintle et al., 2011). The only available selected-response assessment that measures this content area is the GOALS assessment, but as mentioned, this assessment includes very few items that measure statistical literacy. New assessments need to be created to meet the needs of instructors teaching with these newer methods.

Research on Students' Understanding of Statistical Literacy

The studies described in this section were conducted to learn about students' statistical literacy knowledge. Participants in the studies range from students at the elementary-level to adult learners. Content areas that were studied include defining statistical terms, understanding percentages and probabilities, reading graphs and tables, exploring and analyzing data, understanding and interpreting statistical results, and understanding sampling and variability.

Statistical literacy at the elementary and secondary school levels. There are multiple studies that have been conducted with elementary and secondary school students that focus on terminology. Watson and Kelly (2003) conducted a two year longitudinal study with 738 Australian students in Grades 3, 5, 7, and 9. Students were asked to define the words *random*, *sample*, and *variation*. Students in Grades 3 and 5 were only asked to define the word *sample*. Two years later the students were asked to define the words again. Student responses were coded with a score between zero and three. No differences were found in longitudinal performance. Students in Grade 5 had a significantly higher mean score for the definition of the word *sample* than students in Grade 3. Other comparisons of mean scores between the grades were not statistically significant. A subset of students had lessons on chance and variation; they were asked to define the words again. The average scores for those students increased from before to after the lessons for all words except for students in Grade 9 who only improved on the word *random*. Scores were low for all definitions. For students in Grade 9, 30% could not define *sample*, 48% could not define *random*, and 50% could not define *variation*.

A teaching experiment was conducted with ninth grade students in a mathematics course in New Zealand (Sharma et al., 2012). Eight lessons were created to develop students' statistical literacy. Recordings of class sessions, students' written work, student interview responses, and field notes were analyzed qualitatively. The researchers concluded that their students were very capable of doing computations as well as reading tables and graphs.

A study that attempted to categorize 55 fifth and eighth grade students by their ability to read graphs was conducted in Japan by Aoyama and Stephens (2003). Students completed a questionnaire that contained three tasks and they were asked to explain their answers in an interview. Each task included a graph and a set of questions. The first two tasks were created to assess students' basic graph reading ability, and the third task was meant to measure students' ability to create new dimensional information such as suggesting an explanation of a trend shown in a graph. This discussion focuses on the first two tasks because only those tasks measured statistical literacy. Students' responses were categorized as being able to read the data, read between the data, and read beyond the data. For the 17 fifth grade students, 14 were able to reach the third category for both the first and second task. For the 38 eighth grade students, 36 were able to reach the third category on the first task and 32 on the second task. All students in the eighth grade were able to read at least at the second category on both tasks.

Research has been conducted that showed students tend to focus on individual cases rather than the aggregate. For example, Konold, Pollatsek, Well, and Gagnon (1997) interviewed secondary school students who had completed a year-long probability and statistics course and asked them to explore a large dataset. They were asked to use

the dataset to answer research questions. Instead of looking at medians, means, or proportions, students talked about individual cases. Konold et al. also found that students compared counts in group comparisons instead of proportions.

Students between the ages of 11 and 14 in Northern Ireland were provided with multivariate datasets and tested to see if they could explore and understand the data as well as make interpretations of the data (Ridgway et al., 2008). Students used statistics and graphs to explore the data. Classroom observations were conducted to watch students complete activities that required them to use statistics and graphs to explore multivariate data and interpret the data. The researchers claimed that students were able to explore and understand multivariate data; however, students struggled with interpretations.

Jones et al. (2000) found results similar to Ridgway et al. (2008) about students' interpretations. Jones et al. interviewed 20 students in Grades 1 through 5 and asked questions about four categories: describing data, organizing and reducing data, representing data, and analyzing and interpreting data. Students' responses were categorized using the Structure of the Observed Learning Outcome (SOLO) taxonomy (Biggs & Collis, 1982). The four levels were *idiosyncratic*, *transitional*, *quantitative*, and *analytical*. Jones et al. found that students tended to score at higher levels on the describing data questions and worse on the analyzing and interpreting data questions.

Multiple studies have been conducted based on Watson's (2011) three tiers of statistical literacy: understanding of basic statistical problems and terminology, being able to use the basics in the real world, and questioning statistical conclusions and results. In her study of 670 sixth and eighth grade students in Tasmania, Watson (1997) investigated students understanding of pie charts and sampling. Students were asked

questions that were based on media articles and their responses were categorized into the three tiers of statistical literacy. She found that the students did not reach a very high level of statistical literacy; however, eighth grade students did reach a higher level than sixth grade students. She suggested that students need more instruction beyond eighth grade to be able to reach higher levels of statistical literacy.

In a longitudinal study conducted with 38 elementary, middle school, and secondary school students over four years, Watson (2004), as discussed earlier in this paper, described six levels of students' understanding of samples. The first two levels were in Tier 1, the next three levels in Tier 2, and the last level was in Tier 3. Two interviews were conducted with each student three or four years apart and students were categorized into the six levels. She found that 78% of the students moved up by at least one level from the first to second interview. The largest jump was for elementary school children. A majority of the elementary school children went from levels one to three in the first interview up to levels three to six in the second interview. In the last interview, there were still two middle school children who were in the first Tier of statistical literacy. This illustrated that students struggled with the concept of samples.

In a study by Yolcu (2012), the statistical literacy of eighth grade students was measured in six content areas: sample, average, graph, chance, inference, and variation. For each content area, three assessment items were created to assess statistical literacy, one for each of Watson's (2011) three tiers. Selected-response items were based on items from textbooks, curriculum documents, and the existing literature (e.g., Aoyama & Stephens, 2003; Garfield, 2003; Watson & Callingham, 2003). Yolcu found that average

scores were highest for Tier 2 and lowest for Tier 3. Also, students performed better on items about average, inference, and variation as opposed to sample, graph, and chance.

Statistical literacy at the postsecondary and adult level. There were seven studies found in the literature that were conducted with postsecondary students and adults that examined their understanding of statistical literacy. To understand undergraduate students' ability to create, describe, and interpret data presented in tables, a study was designed where 124 students participated in a module in class, and took a two question pretest and a posttest with similar questions (Atkinson et al., 2006). The module taught students basics of statistics such as what a population is and the difference between counts and percentages. Students also learned about time order, causality, reading tables, and making comparisons. In the tests, students were asked to draw conclusions from data presented in tables and explain their answers. Atkinson et al. (2006) found that students were able to understand tables of percentages with very little instruction based on total scores on the posttest. More than half of the students who answered incorrectly on the pretest answered correctly on the posttest.

College students, college teachers, and professional data analysts completed the Statistical Literacy Project international survey to measure their ability to read and compare rates and percentages in tables and graphs (Schield, 2006). A majority of the 37 college teachers and 47 data analysts who participated worked in fields that were considered to be quantitative. For the 85 college students, it was not reported what courses they were taking. For each of the 48 items, the 169 respondents could respond *yes*, *no*, or *I don't know*. Responses were recorded as incorrect for participants who chose *I don't know*. The average error rate for college students was 49%, for data analysts was

44%, and for teachers was 28%. In addition, 2% of college students, 20% of data analysts, and 43% of college teachers got at least 80% correct.

Two studies were conducted to examine statistical literacy knowledge of adults in a medical context. In the first study, adults from the United States ($n = 1009$) and Germany ($n = 1001$) were randomly selected to answer nine probability-related questions (Galesic & Garcia-Retamero, 2010). For the United States participants, the average total percent of items answered correctly was 64.5% and for German participants, the average was 68.5%. Galesic and Garcia-Retamero (2010) concluded that physicians need to take care in presenting statistical information to their patients. In another study, 94 obstetricians and gynecologists answered a questionnaire that contained some questions measuring statistical literacy (Anderson et al., 2014). Specifically, multiple questions examined understanding of probabilities, risks, specificity, and sensitivity. They found that physicians lacked multiple statistical literacy skills. For example, only 66% of physicians correctly answered a question measuring the understanding of sensitivity. However, 89% of physicians correctly answered a question measuring the ability to translate a conditional probability to a frequency.

Students' preconceptions of variability and their understanding gained after taking an introductory statistics course at a community college were studied by using a pretest and posttest (Turegun, 2011). The ARTIST Measures of Spread Topic Scale test, student journals, and interviews were used in the study. Students' scores were significantly higher than what would be expected if they were just guessing ($N = 29$). The average total scores out of 14 items on the pretest and posttest were low, 95% CIs [3.9, 5.2] and [5.6, 7.4] respectively. By comparing students' test scores with their journals, Turegun

found that students underestimated their understanding of variability. In their journals and interviews, students tended to use their own terminology, misusing words such as range and variability.

In a study by Wade (2009), 111 students in four different courses were assessed. The four courses were a statistics course, a research methods course that followed a prerequisite statistics course, a research methods course with no prerequisite statistics course, and a course that appears to not cover research or statistics. A set of 18 selected-response and constructed-response items were chosen from the ARTIST website to measure statistical literacy, reasoning, and thinking. The same items were used on a pretest and posttest. Based on my inspection of the items, ten of the items measured statistical literacy; however, Wade could have used different classifications. She found that students' statistical literacy scores significantly improved for students in the research methods course with a prerequisite statistics course, a non-significant increase for students in the statistics course, and a non-significant decrease for students in the research methods course with no prerequisite statistics course as well as the control group.

Pre-service teachers were the focus of a study conducted by Chick and Pierce (2012). These pre-service teachers were asked to create an appropriate lesson plan for an activity about a dataset and graphs in a study about issues in using real-world data. The activity the pre-service teachers created was for sixth grade students in an advanced math course. Lesson plans included a list of questions to provide to students. There were four kinds of questions: reading data, interpreting data, consideration of context, and attention to implications. Out of 54 pre-service teachers who participated, half were given a preliminary activity and the other half were not. Watson and Callingham's (2003) six

levels of statistical literacy were used to categorize the lesson plans. Pre-service teachers who had a preliminary activity created lessons that contained stronger statistical content. Over half of the lesson plans contained the lowest three levels of statistical literacy and none of the lesson plans reached the highest level of statistical literacy. The average number of questions about interpretations was the lowest of the four kinds of questions, with the highest average being the number of questions about context.

Pierce and Chick (2013) conducted another study with 704 primary and secondary school teachers in Australia. The purpose of the study was to examine teachers' statistically literacy abilities related to boxplots since teachers were routinely shown their students' large-scale assessment results in the form of boxplots. Teachers were asked to answer five selected-response items about an example assessment report that used boxplots. In an additional two constructed-response items, teachers were asked to provide explanations for some of their answers to the selected-response items. There were some selected-response items that teachers did well on; most teachers could compare boxplots to determine a school's poorest area of performance and most teachers understood that the length of the box was a representation of spread. However, the items that asked teachers to explain their answers for the selected-response items that they did well on were mostly incorrect. In addition, there were selected-response items that teachers answered very poorly. For example, only 9% of teachers were able to correctly compare the range of assessment results for a specific school with the states assessment results that were presented in boxplots. Pierce and Chick concluded that teachers are able to read boxplots only at a superficial level.

Summary of studies of students' understanding of statistical literacy. The studies described above cover a wide range of topics. At the elementary and secondary school level, the abilities to write definitions (Watson & Kelly, 2003), understand averages (Yolcu, 2012), read graphs (Aoyama & Stephens, 2003; Sharma et al., 2012; Watson, 2011; Yolcu, 2012), read tables (Sharma et al., 2012), understand sampling (Watson, 2004; Watson, 2011; Yolcu, 2012), explore data (Konold et al., 1997; Ridgway et al., 2008), and analyze data (Jones et al., 2000) were all studied. At the postsecondary and adult level, topics covered included boxplots (Pierce & Chick, 2013), tables (Atkinson et al., 2006), percentages and probabilities (Anderson et al., 2014; Galesic & Garcia-Retamero, 2010; Schield, 2006), variability (Turegun, 2011), data exploration (Chick & Pierce, 2012), and statistical literacy as a whole (Wade, 2009). Also, the ages of participants spanned from students in Grade 1 (Jones et al., 2000) to adults (Anderson et al., 2014; Galesic & Garcia-Retamero, 2010; Schield, 2006).

There were positive and negative results from the studies described above. Students' statistical literacy abilities were better for students in higher grade levels (Aoyama & Stephens, 2003; Watson, 2011; Watson, 2004; Watson & Kelly, 2003). In multiple studies, students had difficulty with interpretations (Jones et al., 2000; Ridgway et al., 2008; Yolcu, 2012). In the study conducted by Chick and Pierce (2012), pre-service teachers did not place a lot of emphasis on interpretations in their lesson plans. In general, most researchers concluded that students and adults did not reach a very high level of statistical literacy (Anderson et al., 2014; Galesic & Garcia-Retamero, 2010; Pierce & Chick, 2013; Schield, 2006; Turegun, 2011; Wade, 2009; Watson, 2011). On the positive side, Ridgway et al. (2008) found evidence that students were able to explore

and understand multivariate data. In addition, some studies found that students could successfully describe data (Jones et al., 2000; Sharma et al., 2012). Lastly, students were able to understand tables of percentages (Atkinson et al., 2006).

Teaching Statistical Literacy

Statistical literacy courses started becoming popular in the late 1970's (e.g., Haack, 1979; Moore & Notz, 1979), however, suggestions about what should be taught in an introductory statistical literacy course go back as far as 1951 (Walker, 1951). Walker suggested that an emphasis should be placed on "concepts, logic, sources of data, interpretation, [and] avoidance of common fallacies" (p. 8). Topics that Walker mentioned are the average, variation, sampling, and interpretations of graphs and tables. Another suggestion was to teach topics that help students become informed citizens (Utts, 2003). Utts suggests that all introductory statistics courses teach seven topics that were commonly misunderstood: conclusions about cause and effect, statistical versus practical significance, no effect versus no statistical effect, bias, coincidences are not uncommon, confusion of the inverse, and understanding that variability is natural. Other suggestions list specific content topics to teach in a statistical literacy course such as data and experimental design, probability, variability, descriptive statistics, and conclusions and inference (Kaplan & Thorpe, 2010).

There are many statistical literacy courses taught which are often referred to as Stat 100, as opposed to Stat 101 which is a more traditional introductory statistics course (Rossman, 2007). A Stat 100 course focuses on statistical literacy and conceptual understanding. There is less emphasis on particular methods and is targeted at consumers of statistics. Rossman believes that a majority of students would benefit more from a Stat

100 course than a Stat 101 course and therefore colleges should offer more Stat 100 courses.

The statistical literacy courses for postsecondary students described in this section included content such as the ones described in the previous paragraphs. In addition, all of the courses were created to teach this content through the media to make statistics relevant to students' daily lives. Articles on four statistical literacy courses were found in the literature and were reviewed in the following section. Most statistical literacy courses use a textbook specifically written to focus on statistical literacy. The two most popular statistical literacy textbooks are also reviewed.

Courses designed to teach statistical literacy. A statistical literacy course was taught to undergraduate students at the University of Kentucky by Haack (1979). Students were not taught how to do any calculation and instead were taught how to interpret results presented in the media. The course also included learning how to detect misuses of statistics that involved more than statistical literacy. In the same year that Haack published an article about his course, the *Statistics: Concepts and Controversies* textbook was published by Moore and Notz (1979). The topics included in the textbook were collecting data, organizing data, and drawing conclusions from data. The textbook is currently in its 8th edition (Moore & Notz, 2013) and includes data production, data analysis, probability, and statistical inference.

The *Chance* course was designed to let students experience probability and statistics through the media (Snell, 1999a, 1999b). The course is focused on statistical literacy; however, it does include a little statistical reasoning and thinking. Topics are presented with current media articles. Students learn what a statistically literate person

should be able to do with an article such as recognize the difference between a good and bad graph. Students will also learn to be more informed, critical readers (Snell, 1999a), and this would be considered to be the statistical thinking portion of the course. In the course, Snell had students read articles in a group and then talk over a couple of discussion questions.

Seeing Through Statistics is a textbook that teaches statistical literacy using the media and journal articles (Utts, 2005). Topics such as the benefits and risks of using statistics, sampling, and confidence intervals are presented through media articles, journal articles, and case studies. Media articles are also traced back to their original publications and students learn statistical thinking skills such as questioning how results are presented in the media. This book was recommended by Rossman (2007) to use in a statistical literacy course. Rossman claimed that a statistical literacy course should have a focus on conceptual understanding, with less emphasis on specific methods. The course should be designed for consumers and not producers of statistical analysis. According to Rossman, *Seeing Through Statistics* focuses on conceptual understanding and is intended for consumers.

Lies, Damned Lies, and Statistics is a statistical literacy course taught at the University of Auckland (Budgett & Pfannkuch, 2007). The course topics include media reports, surveys and polls, experimentations, risk, statistical reasoning, and statistics and the law. Compared with statistical literacy courses based on the media, as discussed above, this course has a larger focus on statistical thinking because of its large emphasis on critical evaluation. The course also has a narrower audience which includes “aspiring

journalists, politicians, sociologists, lawyers, health personnel, business people, and scientists” (p. 1).

A statistical literacy course was created by Schield (2004) that focuses on chance, inference, and confounding. Specific topics include conditional probability using ordinary language, measuring associations, confidence intervals, and a large emphasis on confounding. The topics were chosen based on what students see in everyday life, including the media. Students learn critical thinking skills; therefore, the course includes more statistical thinking than statistical literacy.

Summary of courses designed to teach statistical literacy. Because statistical literacy includes being able to read statistical information, it makes sense that most statistical literacy courses include opportunities for students to learn statistics by examining what is reported in the media. There are many examples in the media where statistical findings are poorly written or written in a possibly deceiving way (Crossen, 1996; Gal & Murray, 2011). As a result, it appears that instructors of statistical literacy courses want to teach students to be critical thinkers. This could explain why many statistical literacy courses include aspects of statistical reasoning and statistical thinking.

Changes in Introductory Statistics: Simulation-based Methods

How statistical literacy is taught and what content is covered in an introductory statistics course has changed immensely over the years (Chance & Garfield, 2002; Cobb, 1992; Moore, 1997; Utts, 2003). Some introductory statistics courses are moving away from the *consensus curriculum* (Cobb, 2007). The consensus curriculum emphasizes aspects of statistical literacy such as understanding terminology and computations. This curriculum includes multiple topics used to understand and perform normal

approximations such as the central limit theorem and conditions necessary to use normal approximations. According to delMas et al. (2007), “Partially in response to the difficulties students have with learning and understanding statistics, a reform movement was initiated in the early 1990s to transform the teaching of statistics at the introductory level” (p. 29). One reason for this shift was instructors wanted students to be able to apply statistics in their daily lives. Chance (1997) said that “Many statistics courses are shifting focus, emphasizing skills such as the ability to interpret, evaluate, and apply statistical ideas over procedural calculations” (para. 2).

Another reason for the shift away from the consensus curriculum was the availability of technology which allows students to automate the computations. Technology is being used often to perform computations for students, and this allows students to work with larger, real datasets (Moore, 1997). Technology makes it possible to run different kinds of statistical analyses. For example, it was not practical to teach resampling methods until proper technological tools were available. Resampling Stats was a revolutionary computer language that made resampling methods accessible to students as well as adult workers (Simon & Bruce, 1990). As a result, it is now possible to teach resampling methods in introductory statistics courses. Resampling methods can be taught in addition to normal-based tests such as the t -test (Tittle et al., 2011) and some instructors have eliminated normal-based tests from their courses (Garfield et al., 2012).

Description of simulation-based methods. There are many resampling methods that have been taught in introductory statistics courses. Three commonly used methods are randomization tests, permutation tests, and bootstrap confidence intervals (Rodgers, 1999). Randomization tests that compare two groups were first described by Fisher

(1936). In a randomization test, the null model is simulated by combining two groups of data and randomly assigning observations to the two groups by sampling without replacement in order to create randomization samples. Many randomization samples are taken and the observed statistic is compared with the distribution of randomization statistics in order to determine if the results are significant. The terms *randomization tests* and *permutation tests* are often used interchangeably because the computations are the same; however, according to Ernst (2004), they have different underlying models. Randomization tests are used to compare groups in a randomized experiment and a permutation test is used in an observational study that included random sampling. For simplicity, in the remainder of this article, both of these types of tests will be referred to as randomization tests.

Bootstrap distributions differ from randomization distributions because the bootstrap samples are drawn with replacement (Efron, 1979). To create the bootstrap distribution, the observed sample is considered to be representative of the population, and therefore, bootstrap samples can be taken from the observed sample to estimate a population parameter (Johnson, 2001). Bootstrap confidence intervals can be found using different methods. Two methods that have been used in an introductory statistics course are looking at the percentiles in a bootstrap distribution and using the standard error of the bootstrap distribution (Engel, 2010).

Rationale for teaching simulation-based methods in an introductory statistics course. There are multiple reasons why simulation-based methods have been recommended to be taught in an introductory statistics course. Simulation-based methods have been argued to be simple and easily grasped (Cobb, 2007; Simon & Bruce, 1990),

accessible to students without mathematical backgrounds (Simon, Atkinson, & Shevokas, 1976; Simon & Bruce, 1990), intuitive (Engel, 2010; Hesterberg, 1998), require minimal prior knowledge (Chaput, Girard, & Henry, 2011; Wood, 2004), and can be taught on Day 1 (Rossman, 2007). Therefore, more time is available for other topics (Cobb, 2007). In addition, simulation-based methods have been said to be visual (Engel, 2010), promote understanding (Hesterberg et al., 2003), and promote cooperative learning (Romeu, 1995). The model matches the data collection method and mimics a real-world problem, and this has been claimed to help students understand inferential methods (Cobb, 2007; Hesterberg, 1998; Rossman, 2007). Mimicking the data collection method emphasizes the scope of conclusions that can be made (Rossman, 2007). If normal-based methods are taught in addition to simulation-based methods, students are expected to gain better understanding of the theoretical solutions (Simon & Bruce, 1990; Tanis, 1992).

There are many methodological reasons why simulation-based methods have been advocated to be taught instead of normal-based methods. There are fewer assumptions associated with simulation-based methods than normal-based methods (Cobb, 2007; Ernst, 2004; Hesterberg et al., 2003; Simon & Bruce, 1990). As an example, in the beginning of a course, students are taught that the median is resistant to outliers, which is often ignored when conducting hypothesis tests for a mean (Hesterberg, 1998). Teaching simulation-based methods can address that concern because simulation-based methods can be used for many statistics, including the median (Engel, 2010; Hesterberg, 1998; Hesterberg et al., 2003; Rossman, 2007; Wood, 2004). There are situations where simulation-based methods are more accurate than normal-based methods (Hesterberg et

al., 2003), and teaching simulation-based methods has been claimed to be truer to Fisher's vision (Cobb, 2007; Ernst, 2004; Rossman, 2007).

Introductory statistics courses and textbooks that teach simulation-based methods. There are a variety of introductory statistics courses and textbooks that cover simulation-based methods. These courses and textbooks could include simulation-based methods as an additional topic at the end of a course (Hesterberg et al., 2003), at the beginning of the course followed by normal-based methods (Lock, Lock, Lock Morgan, Lock, & Lock, 2013), throughout the course alongside normal-based methods (Tintle et al., 2013), or as the only inferential method in the course (Garfield et al., 2012). Each of these types of courses and textbooks are described in this subsection.

Bootstrap Methods and Permutation Tests (Hesterberg et al., 2003) is a companion chapter for *The Practice of Business Statistics* textbook (Moore, McCabe, Duckworth, & Sclove, 2003). This chapter could be used as an add-on to the consensus curriculum. The chapter begins by covering the bootstrap distribution. Bootstrap t and bootstrap percentile confidence intervals for proportions, means, trimmed mean, median, and correlation are included. One-sample and two-sample situations are covered. The chapter covers the accuracy of bootstrap confidence intervals, the bootstrap for a scatterplot smoother, and the bootstrap bias-corrected accelerated and bootstrap tilting methods. The chapter also describes permutation tests for two-sample and matched-paired situations. Advantages and disadvantages of these methods are discussed. The software that is described for this chapter is S-PLUSTM (S-Plus, 2007). A majority of the examples provided in the chapter are not based on real studies.

Statistics: Unlocking the Power of Data is a textbook that teaches simulation-based methods and then follows up with the normal-based methods (Lock et al., 2013). The book contains four units: Data, Understanding Inference, Inference with Normal and t -Distributions, and Inference for Multiple Parameters. The first unit is similar to the beginning of a consensus curriculum textbook and covers collecting and describing data. The second unit starts by describing simulated sampling distributions and then proceeds to cover, in two sections, bootstrap confidence intervals for means and proportions in one-sample, two-sample, and matched-pairs situations. The first section describes computing bootstrap confidence intervals using standard errors and the second section describes the percentile approach. Bootstrap confidence intervals are followed by randomization tests. Unit 3 switches over to the consensus curriculum but continues to introduce each new topic by relating back to the simulation-based methods. The software that accompanies the book, StatKey, was created by the textbook authors. The book includes real-world applications and datasets with multiple variables.

The *Concepts of Statistical Inference* (CSI; Holcomb et al., 2011; Rossman & Chance, 2008) course follows a similar content outline as the Lock et al. (2013) textbook. Simulation-based methods are presented in the form of modules and can be used in addition to a consensus curriculum textbook. The entire statistical process is stressed from the research question to the scope of conclusions, not just the analysis. Activities emphasize the entire statistical process and are based on research studies and classroom studies. Applets are used for the simulation-based analysis. Tintle et al. (2011) created activities based on the CSI modules and are working with the CSI team to write a textbook including simulation-based and normal-based methods. The new course and

textbook are called *Introduction to Statistical Investigations* (ISI; Tintle et al., 2013). Inference is covered early in the book and descriptive statistics are introduced as needed throughout the book. Hypothesis testing is introduced using a simulation-based approach and is immediately followed by the corresponding normal-based method. For example, the first inferential technique in the textbook is a one-sample proportion hypothesis test which is first taught via simulations in an applet and then using the normal approximation.

The *Change Agents for Teaching and Learning Statistics* (CATALST; Garfield et al., 2012) curriculum is different from the previously discussed curricula because of the emphasis on modeling. This course removed all normal-based methods from the curriculum. Students are exposed to ideas of inference on the first day of class and are encouraged to create their own models (Ziegler & Garfield, 2013). The CATALST course has three units: Chance Models and Simulation, Models for Comparing Groups, and Estimating Models using Data. In the Chance Models and Simulation unit, students explore and learn how to model randomness and informally conduct a simulation-based hypothesis test for one sample. The Models for Comparing Groups unit focuses on randomization tests and the Estimating Models using Data unit focuses on bootstrap confidence intervals. The course is activity-based and fosters cooperative learning. Activities are based on research studies and stress answering a research question and the scope of conclusions. The software students use is TinkerPlots™ (Konold & Miller, 2011), which allows students to explore and create their own models.

Evaluating students' understanding of statistics in a simulation-based introductory statistics course. As simulation-based methods are introduced in the

introductory statistics course, it is important to find out what statistical concepts students understand as a result of using these methods. Seven studies were located and evaluated. These studies involved undergraduate students who were enrolled in a simulation-based course. The first study was about students who took a course that taught the Monte Carlo simulation method in addition to normal-based methods and the other studies were about students who took the ISI course, CSI course, and CATALST course.

Three studies were conducted by Simon et al. (1976) to test undergraduate students' understanding of statistics upon completion of an introductory statistics course that incorporated normal-based methods and the Monte Carlo simulation method. Each study was conducted at a different institution and used different final exams to assess students' understanding of statistics. The first study included 25 students who were mostly economic and business majors. The final exam for the course included four constructed-response questions where students could choose to use normal-based methods or the Monte Carlo method. More than half of the responses were completed using the Monte Carlo method, and students who used the Monte Carlo method were more likely to get the answer correct. The second study compared three curricula taught in a general mathematics course: one using the consensus curriculum, one including the Monte Carlo method with computers, and the other including the Monte Carlo methods without computers. Enrollments in the sections were 19, 39, and 13, respectively. The average scores on a seven question exam were significantly higher for both of the Monte Carlo classes compared with the average scores for the consensus curriculum course. The last study also compared courses that used the consensus curriculum and the Monte Carlo curriculum. In two semesters, 58 students enrolled in the consensus curriculum sections

and 55 in the Monte Carlo sections. Students in this study were not mathematically inclined. In the first semester, students had a higher average final exam score in the consensus curriculum section but in the second semester, students had a higher average in the Monte Carlo section.

The statistical knowledge that college students gain in the ISI course and a consensus curriculum introductory statistics course was investigated by comparing scores on a pretest and a posttest (Swanson, Tintle, VanderStoep, Holmes, & Quisenberry, 2010; Tintle, et al., 2011). The CAOS test was administered as a pretest and a posttest to 195 students who took the consensus curriculum course and 202 students who took the ISI course. The average scores for students in the two curricula were compared with each other as well as with the national sample of students who have taken the CAOS test. Students who took the ISI course had a higher gain in statistical knowledge than the consensus curriculum and the national sample, especially for items about statistical inference. A subset of the students, 79 from the consensus curriculum course and 76 from the ISI course, were followed up four months after the course ended and took the CAOS test a third time (Tintle, Topliff, Vanderstoep, Holmes, & Swanson, 2012). The researchers found that the students who took the ISI course retained more conceptually than students who were exposed to the consensus curriculum, particularly in statistical inference. The results suggest that because the ISI curriculum emphasizes statistical inference earlier in the course, students are more likely to remember conceptual knowledge of statistics.

A study that compared undergraduate students who took an introductory statistics course using consensus curriculum to students in a CSI course examined students

understanding of hypothesis tests (Holcomb et al., 2011). The final exam included five items from the CAOS test that measured students' comprehension of hypothesis testing. The percentages correct for each item were compared for the two classes as well as to students in the study conducted by Tintle et al. (2011) and to the national sample of students that had taken the CAOS test. The results for the simulation-based courses were similar to each other. Four out of the five items had a higher percentage correct for the simulation-based courses than for the consensus curriculum courses and the national sample. The results suggest that students develop a better understanding of statistical significance and p -values in a course using the simulation-based curriculum than students taught with the consensus curriculum.

A 3-month teaching experiment was conducted with 102 students taking the CATALST course (Garfield et al., 2012). Students completed two assessments of statistical knowledge at the end of the semester: the GOALS assessment and the Models of Statistical Thinking (MOST) assessment. As mentioned previously, a majority of the selected-response items on the GOALS assessment were designed to measure statistical reasoning. Sixteen items on the GOALS assessment were identical or similar to items on the CAOS test, which allowed the researchers to compare the students in the teaching experiment to the national sample of students that took the CAOS assessment. The MOST assessment included constructed-response item designed to measure statistical thinking. On the GOALS assessment, students answered 66% (SD = 12.3%) of the items correctly that measured basic statistical literacy and reasoning. Specifically, students performed very well on items involving graphical representations of data and items about comparing observed results a null-distribution from a randomization test. However,

students did not perform as well on items asking about interpretations of p -values. Based on the comparison of the GOALS and CAOS results, the researchers concluded that students who took the CATALST course have the same or better statistical literacy and reasoning skills than students who took other introductory statistics courses. Based on the results from the MOST assessment, there was evidence that many students were beginning to think statistically.

Summary of simulation-based introductory statistics courses. Changes in the teaching of statistics in an introductory course include teaching simulation-based methods. Simulation-based methods that have been taught in introductory statistics courses include randomization tests and bootstrap confidence intervals (Rodgers, 1999). Many reasons have been provided about why simulation-based methods should be taught in an introductory statistics course, and these include that they are easily grasped (Cobb, 2007; Simon & Bruce, 1990) and can be taught on the first day of class (Rossman, 2007).

The courses and textbooks described in this section cover both simulation-based and normal-based methods except for the CATALST course (Garfield et al., 2012) which includes only simulation-based methods. The *Bootstrap Methods and Permutation Tests* companion chapter written by Hesterberg et al. (2003) is used at the end of a consensus curriculum course. The Lock et al. (2013) book teaches simulation-based methods followed by normal-based methods and the ISI (Tintle et al., 2013) course alternates between simulation-based and normal-based methods. See Table 5 for a comparison of the simulation-based courses and textbooks.

Table 5

Courses that Teach Simulation-based Methods

Course	Inferential methods in the order that they appear	Textbook	Software
The Practice of Business Statistics (Hesterberg, et al., 2003)	Consensus curriculum methods, bootstrap confidence intervals, and permutation tests	<i>The Practice of Business Statistics</i> (Moore, et al., 2003)	S-PLUS™
Lock (Lock et al., 2013)	Randomization tests, bootstrap confidence intervals, and normal-based methods	<i>Statistics: Unlocking the Power of Data</i> (Lock et al., 2013)	StatKey
ISI (Tintle et al., 2013)	Simulation-based methods and normal-based methods alternate throughout the curriculum	<i>Introduction to Statistical Investigations</i> (Tintle et al., 2013)	Applets
CATALST (Garfield et al., 2012)	Modeling, randomization tests, and bootstrap confidence intervals	<i>Statistical Thinking: A Simulation Approach to Modeling Uncertainty</i> (Catalysts for Change, 2013)	TinkerPlots™

The focus for each of the courses and textbooks described in this section are different. Hesterberg et al. (2003) emphasizes mechanics and uses S-PLUS™ which is used by statisticians. In contrast, Lock et al. (2013) focuses on data analysis and uses StatKey, which was created for educational purposes, allowing students to visualize distributions. The statistical process is at the core of the ISI course (Tintle et al., 2013) and the applets incorporated are specific to particular activities. The underlying theme for the CATALST course (Garfield et al., 2012) is modeling, and therefore TinkerPlots™ software is used because it requires students to create their own models. The activities created for the ISI and CATALST courses are based on real research studies and highlight the research question and scope of conclusions. The Lock et al. (2013) textbook uses real data and real applications.

Overall, the evidence to date suggests that students perform better in introductory statistics courses that include simulation-based methods. Each study described in this section about students' understanding of statistics in a simulation-based introductory course observed certain results favoring the consensus curriculum courses, but most of the results favored the simulation-based courses. Multiple studies have found that students understand inferential methods better in a simulation-based course (Holcomb et al., 2011; Swanson et al., 2010; Tintle et al., 2012; Tintle et al., 2011).

Discussion of the Literature Reviewed

The purpose of this critical review of the literature was to try to understand and describe statistical literacy, how it has been taught and assessed in introductory college-level statistics courses, and to re-conceptualize what statistical literacy means in the context of new curriculum based on simulation methods. A working definition of

statistical literacy was created to guide the review of the literature. Assessments of statistical literacy, studies about students' understanding of statistical literacy, and courses that teach statistical literacy have been summarized. In this section, a synthesis and critique of the literature that was described in the previous sections are provided, and a case is made for new assessments and studies about students' understanding of statistical literacy.

Summary and critique of the literature reviewed. Definitions of statistical literacy vary from knowing the basic language of statistics (Garfield et al., 2002) to communicating, interpreting, and being critical of statistical information (Gal, 2002). The definition of general literacy is the ability to read and write ("Literacy," n.d.a; "Literacy," n.d.b) and was used to make a case for defining statistical literacy as the ability to read, understand, and communicate statistical information. Building on this definition, statistical literacy includes understanding definitions and terms, creating graphs, and doing computations. This working definition was problematic because it was not clear whether interpretation was a part of statistical literacy or statistical reasoning.

Given the issue of whether interpretation was a part of statistical literacy or statistical reasoning, a clarification in the definitions put forward by Garfield and delMas' (2010) was suggested. Graphs and descriptive statistics are often reported in the media and therefore creating and interpreting graphs and descriptive statistics are basic skills that citizens should have. Thus, it is proposed that making an interpretation of graphs and descriptive statistics will be considered statistical literacy.

Making interpretations of confidence intervals and hypothesis tests appear to be a part of statistical reasoning because general reasoning has been defined as drawing

inferences or conclusions (“Reasoning,” n.d.). Furthermore, in order to interpret inferential statistics, students need to make appropriate connections such as connecting the interpretation to the data collection method. For example, to make a correct interpretation about an observational study, students need to conduct an analysis, connect the design of the study to the interpretation, and determine that cause and effect statements cannot be made because there was no random assignment.

The large-scale assessments reviewed in this paper measure both statistical literacy and statistical reasoning. The assessments that were created to measure outcomes for college students enrolled in an introductory statistic course appear to do a good job of measuring statistical reasoning in a consensus curriculum, but a majority of the assessments do not have many items that measure statistical literacy. In addition, those assessments have items that are not relevant to students who are enrolled in simulation-based courses. For example, the ARTIST Topic Scale item in Figure 5 is meant to test whether or not students understand how a median is computed. Simulation-based courses focus on inference and students are not necessarily taught how to compute a median. The focus of the GOALS assessment was to measure statistical reasoning for students enrolled in simulation-based courses, however, there are a few items that measure statistical literacy. The GOALS assessment is a start, but new assessments are needed to assess statistical literacy in a simulation-based course.

4. You give a test to 100 students and determine the median score. After grading the test, you realize that the 10 students with the highest scores did exceptionally well. You decide to award these 10 students a bonus of 5 more points. The median of the new score distribution will be _____ that of the original score distribution.
- lower than
 - equal to
 - higher than
 - depending on skewness, higher or lower than

Figure 5. Example item on the ARTIST Measures of Center Topic Scale test that measures statistical reasoning.

The research studies discussed here involved college-level students taking an introductory statistics course taught using the consensus curriculum. A wide range of topics was researched, including the ability to make interpretations of data in tables, read and compare rates in tables and graphs, and understand variability. Wade (2009) focused on the bigger picture by not focusing on a specific content topic; instead, she attempted to measure statistical literacy, reasoning, and thinking. Her study was problematic because even though she provided the assessment items that were administered to students, she did not say which items were used to assess each of the three outcomes. Therefore, it is not clear if the results for the three outcomes that Wade provided actually represent the three outcomes as defined in this paper. In addition, the assessments used in these studies could have been better. The only studies that used a pretest and posttest for students who completed an introductory statistics course were Atkinson et al. (2006), Turegun (2011), and Wade (2009). Only one study used a high quality assessment (ARTIST Topic Scale) that was created for research purposes (Turegun, 2011). Yolcu (2012) used the existing literature to choose assessment items and Wade (2009) used the test bank available on the ARTIST website. There was no validity or reliability evidence for any of the assessments

used. Overall, these studies answer some good questions, but they are just scratching the surface.

This review also examined statistical literacy courses and textbooks. A majority of the statistical literacy courses and textbooks described in this paper were created to teach statistics through the media and make statistics applicable to students' daily lives. Furthermore, these courses and textbooks do not teach only statistical literacy as defined in this paper. Most of these courses also include aspects of statistical thinking, such as critical thinking. Therefore, it appears that these courses and textbooks do a good job of teaching statistics through the media, but saying they are statistical literacy courses and textbooks is deceiving because they teach other outcomes, such as statistical thinking.

In light of changes in statistics education, simulation-based methods are being taught in college-level introductory statistics courses. The simulation-based courses and textbooks described in this paper appear to be well thought-out, and research has been done to assess students' knowledge in some of these courses. These simulation-based courses and textbooks emphasize statistical reasoning and thinking, but an argument could be made that students could be statistically literate, to some extent, after completing a simulation-based course. Students are not explicitly taught statistical literacy as a goal in and of itself; however, while students learn how to reason and think statistically, they also gain basic statistical knowledge. For example, being able to create and interpret a graph from a sample of data is a statistical literacy skill that is not taught as an activity in the CATALST course. However, students do complete activities to learn how to create and interpret results from a bootstrap distribution. If students understand

how to create and interpret graphs of a bootstrap distribution, it is plausible that students could create and interpret graphs of data from a sample without further instruction.

Multiple research studies found that students understood inferential methods better in a simulation-based course compared to a course taught with the consensus curriculum (Holcomb et al., 2011; Swanson et al., 2010; Tintle et al., 2012; Tintle et al., 2011). All of the research studies about students in simulation-based courses described in this paper compared students in a simulation-based course to students in a consensus course. This is helpful for instructors who are considering teaching simulation-based methods in their introductory statistics courses. However, these studies do not help instructors decide how to incorporate simulation-based methods into their courses. The simulation-based courses and textbooks described in this review have different approaches to teaching simulation-based methods, but the research studies do not provide evidence for which methods are most beneficial to students.

Introductory statistics courses taught using the consensus curriculum, statistical literacy courses, and simulation-based courses have multiple differences. For example, the focus on statistical literacy, reasoning, and thinking differ. Often, the consensus curriculum has a large focus on statistical literacy, statistical literacy courses teach all three outcomes, and simulation-based courses focus on statistical reasoning and thinking. In addition, students in simulation-based courses and statistical literacy courses experience statistics in real-world situations; however, the types of real-world situations could differ. The statistical literacy courses focus on teaching using the media (e.g., Snell, 1999a, 1999b) and the simulation-based courses focus on teaching using research articles. There are some exceptions, however; the statistical literacy textbook *Seeing Through*

Statistics (Utts, 2005) uses research articles and ties media articles back to their original research articles, and the simulation-based textbook *Statistics: Unlocking the Power of Data* (Lock et al, 2013) refers to media articles. Students in consensus curriculum courses can also be exposed to real-world situations but that is not always practiced. The components of statistical literacy included in the courses are different because of the various differences in these two types of courses.

I propose that the main components of statistical literacy in a simulation-based course include understanding terms, simulations, and inferential techniques. Statistical literacy in a simulation-based course can include statistical literacy topics included in the consensus curriculum, such as computing descriptive statistics and interpreting graphs, but those skills are not the end goal in a simulation-based course.

Formulation of the problem statement. This literature review aimed to examine many aspects of statistical literacy in order to redefine this learning outcome in light of changes to the introductory statistics course, namely the move to include simulation-based methods. As more instructors begin to incorporate a simulation-based curriculum, they may want to know how statistically literate their students are in the new curriculum. In order to determine how statistically literate students are in a course including simulation-based methods, new assessments need to be created. A new statistical literacy assessment should include real-world problems and measure the important outcomes in a simulation-based course. The assessment could also include selected-response items so that scoring is easier which may be important for research.

A new statistical literacy assessment could be used for a variety of purposes. An assessment could be used as a pretest to determine what statistical literacy skills students

have prior to taking an introductory statistics course at the postsecondary level. Statistics instructors could use an assessment in their classes to see which statistical literacy topics students understand and which topics they do not understand. Researchers could also use an assessment to determine how much statistical literacy students gain in an introductory statistics course. Also, researchers could use a new assessment to determine if the amount of statistical literacy gained in introductory statistics courses taught with varying teaching methods or curricula is different. For example, are students more statistically literate when they are taught simulation-based methods first and then normal-based methods, when they alternate back and forth between simulation-based and normal-based methods, or when only simulation-based methods are included in the course? To address these questions, new assessments of statistical literacy are needed. In light of changes to the introductory statistics course, namely the move to include simulation-based methods, the topics of statistical literacy that should be assessed need to be considered important to instructors who teach using different methods.

Chapter 3

Methods

The literature review in the previous chapter showed the need for new assessments. In this chapter, the development of the Basic Literacy In Statistics (BLIS) assessment is described. The BLIS assessment was designed to determine what statistical literacy knowledge students have in an introductory statistics course, at the postsecondary level. Since the intent is to use this instrument as an assessment in a more modern version of the introductory statistics course, the instrument includes items on simulation-based methods as well as some of the more traditional topics.

This chapter starts by including a description of what reliability and validity are followed by a description of the study. A short overview of the study is then provided, as well as a description of the development process. The development process included the creation of the test blueprint and assessment, the collection of reliability and validity evidence, and the evaluation of the final assessment. Following the description of the development process, details regarding the data analyses are presented. The chapter ends with a summary of the methods described.

Determining the Quality of an Assessment

According to AERA, APA, and NCME (1999), reliability and validity are two characteristics of assessments that need to be examined. Definitions of these terms and how to examine them vary. In this section, definitions of reliability and validity, as well as sources of low reliability and low validity, are described. Furthermore, methods to determine the extent to which an assessment is reliable and valid are discussed.

Reliability. The reliability of an assessment refers to the consistency of measurements made from an individual (AERA, APA, & NCME, 1999; Thorndike & Thorndike-Christ, 2010). In other words, if an assessment is administered to the same individuals multiple times, the results will be similar each time (Weathington, Cunningham, & Pittenger, 2010). Low reliability can be the result of measurement error. All assessments have some measurement error due to natural variability, but there are some sources of measurement error that can be minimized: instrument error, participant variability, researcher variability, and environmental variability (Weathington et al., 2010). Instrument error includes wording and organizational issues, participant variability includes fatigue and misunderstanding of items, researcher variability includes errors in recording, and environmental variability includes distractions and differences in testing locations. Evidence of reliability as it relates to sources of measurement error can be collected throughout the assessment development process. According to Buckendahl and Plake (2006), without examining the reliability evidence for an assessment, other evidence of validity may not be meaningful.

Validity. The definition of validity is not as clear as the definition of reliability because it refers to the interpretations of test scores and not the assessment itself (Weathington et al., 2010). Interpretations are affected by how much variability there is in participant responses, and therefore reliability is needed in order to have validity. There are two competing views on validity. According to Thorndike and Thorndike-Christ (2010), “The traditional-narrower-view of construct validity involves providing evidence to support assertions that test scores or criterion measures actually tap into the underlying constructs they claim to represent” (p. 186). In contrast, a unified view of

validity is described in the definition specified by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999); “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9).

The unified view takes the perspective that there are not distinct types of validity (AERA, APA, & NCME, 1999) and focuses on validity as related to interpretations of test scores and uses of test scores for applied purposes instead of the test itself (Kane, 1990; Messick, 1989). In order to assess validity, the type of inferences to be made and the purpose of the test must be determined (Thorndike & Thorndike-Christ, 2010). If more than one type of inference is of interest, each should have a separate validity argument. For example, inferences of interest would be different for a pretest and a posttest. Multiple pieces of evidence must be gathered and collated into one validity argument for the reasonableness of inferences. The validity argument must continually be re-evaluated; it is not all or nothing and is a never-ending process.

For the analysis of the proposed assessment, the unified view of validity was used for multiple reasons. First, with both the traditional and unified views, the same analyses would be conducted, but interpretation would be different. Second, the traditional view is segmented and the unified view is holistic. Third, the unified view is promoted by the *Standards* (AERA, APA, & NCME, 1999), which, according to Lin (2006), “are widely recognized as the most authoritative statement of professional consensus regarding the development and evaluation of educational and psychological tests” (p. 27).

Overview of the Study

The purpose of this study was to create an assessment of statistical literacy that can be used to determine what statistical literacy skills students have in an introductory statistics course at the postsecondary level that includes some amount of simulation-based methods in the curriculum. The content included in the assessment must be relevant to a variety of introductory statistics courses: courses that include only simulation-based methods, courses that emphasize simulation-based methods and parametric methods, and courses that focus on parametric methods but include simulation-based methods to help students gain a conceptual understanding of the content. A mixed-methods approach was used to collect evidence of reliability, validity, and value for the BLIS assessment.

The development of BLIS occurred in five phases, and evidence of the value, reliability, and validity was collected throughout the phases. All materials that were developed were reviewed by two faculty members at the University of Minnesota, Dr. Joan Garfield and Dr. Michelle Everson. The faculty members' feedback was used to make edits to the materials.

The test blueprint was developed in the first phase. The preliminary test blueprint consisted of statistical literacy topics compiled from textbooks utilizing simulation-based methods. A review was conducted with six experts in the field of statistics education to provide evidence that the learning outcomes specified in the preliminary test blueprint capture the intended construct: statistical literacy. The reviews were used to modify the test blueprint.

In the remaining phases, multiple versions of the assessment were developed. The preliminary assessment was created using existing selected-response items and newly developed constructed-response items. An expert review of the preliminary assessment was conducted to create BLIS-1. Cognitive interviews, conducted with students, were used to make additional modifications for the next version of the assessment, BLIS-2. This second version of the assessment was administered to students in three different introductory statistics courses as a pilot. The pilot results were used to convert the constructed-response items to selected-response items, which then became BLIS-3. A field test was conducted in which students in multiple introductory statistics courses completed the BLIS-3, and their instructors were surveyed to examine the value of the assessment. Multiple statistical analyses were conducted with the data from the field test to gather further evidence of validity and reliability.

Table 6 presents an outline of the assessment development process. Details of each step of the development process, data collection, and analysis are provided in the following subsections.

Table 6

Overview of Assessment Development, Data Collection, and Analysis

Phase and timeline	Product and analysis	Data collected
Phase I		
September - October 2013	Preliminary test blueprint	Review of textbooks
October - November 2013		Expert review of preliminary test blueprint
November 2013	Evaluate expert review data for preliminary test blueprint	
November 2013	BLIS Test Blueprint-1	
Phase II		
October - November 2013	Item pool Preliminary assessment	
December 2013	Evaluate expert review data for preliminary assessment	Expert review of preliminary assessment
Phase III		
December 2013 - January 2014	BLIS-1	
January - February 2014		Cognitive interviews with students
February 2014	Evaluate students' cognitive interview data	
Phase IV		
February - March 2014	BLIS-2	
March 2014	Evaluate students' responses from pilot	Pilot test
Phase V		
March - April 2014	BLIS-3	
April 2014	Instructor survey	Field test
May 2014	Factor analysis, reliability analysis, analysis based on item response theory	

Test Blueprint Development

Prior to developing an assessment, a test blueprint should be developed (Garfield et al., 2010). A test blueprint is used to outline a set of topics that will be included on an assessment and how many items will be written for each topic (Downing, 2006b).

Assessments should include topics that are deemed to be important so students are not misled about what the key topics are (Garfield & delMas, 2010). Therefore, it is important to carefully choose which topics are included on a test blueprint.

Development of the preliminary test blueprint. In order to create the preliminary test blueprint for the BLIS assessment, a list of statistical literacy topics was first created. The topics of statistical literacy that were included were chosen to be of interest to instructors who incorporate simulation-based methods in their introductory statistics course. The list of topics was created based on a review of introductory statistics textbooks (Gould & Ryan, 2013; Catalysts for Change, 2013; Lock et al., 2013; Tintle et al., 2013) that incorporate simulation-based methods. Topics that were chosen were largely emphasized in all textbooks reviewed and were not specific to one particular textbook. In addition, topics were only included that related to the goal of the assessment, which was to assess students' statistical literacy abilities in a simulation-based introductory statistics course. In light of the goal of the instrument, topics that appear to be more important (e.g., hypothesis tests) were included and topics that are less important (e.g., probability rules) were not included. Topics were considered to be less important if they were not emphasized in two or more of the four textbooks that were reviewed.

In order to determine how many items to write for each topic, statistical literacy learning outcomes were specified for each topic. Each learning outcome corresponded

with one item. Multiple learning outcomes were related to being able to identify, describe, translate, interpret, read, and compute, which are words that have been associated with items measuring statistical literacy (Garfield et al., 2010). A majority of the learning outcomes focused on being able to describe and interpret. For example, one topic is *descriptive statistics*. An appropriate learning outcome would be that a student has the “ability to interpret a standard deviation in the context of data.” Some learning outcomes were written based on examining learning goals and objectives presented in textbooks (Gould & Ryan, 2013; Catalysts for Change, 2013; Lock et al., 2013) and some were taken from the CAOS test blueprint (Joan Garfield, personal communication, September 9, 2013). See Table A1 in Appendix A for a copy of the preliminary test blueprint.

Expert review of the preliminary test blueprint. After the preliminary test blueprint was created, statistics educators were invited to review the preliminary test blueprint and assessment in order to get perspectives from individuals with different expertise. The purpose of the review was to provide evidence of construct validity. Prospective reviewers were provided with an invitation letter, an example of what it looks like for a student to be statistically literate, and the preliminary test blueprint review form in an email. The following sub-sections include details about the participants and procedures for the review.

Participants. Reviews were requested from statistics education researchers, statisticians, statistics instructors who teach introductory statistics courses using simulation-based methods, introductory statistics textbook authors, and statistics assessment experts. Multiple rounds of requests were sent out in October 2013 until six

reviewers were found. Researchers 1 and 2 were recruited from the statistics education program at the University of Minnesota, reviewers 3 and 4 were statisticians who teach introductory statistics using simulation-based methods, and reviewers 5 and 6 were experts in statistics assessment development. All of the reviewers were authors of textbooks for introductory statistics students, and four of the reviewers emphasized simulation-based methods in their textbooks.

Review procedures. An invitation letter was written to invite expert reviewers to participate, which was loosely based on a reviewer invitation letter written for a different assessment study (Park, 2012). The invitation letter included the definition of statistical literacy that was used to create the BLIS assessment, a description of the purpose of the assessment, the purpose of the review, and a deadline of two weeks for the review. See Appendix B1 for the invitation letter that was sent.

Two examples of what it looks like for students to be statistically literate were shared with the reviewers (see Appendix B2). The purpose of creating and providing these examples was to help the reviewers understand the definition of statistical literacy that was used to create the test blueprint. The examples were also meant to guide the reviewers as they decided which learning outcomes would be statistical literacy outcomes. One example provided to reviewers was based on an example from the Gould and Ryan (2013) textbook, and the other was based on an activity from the Catalysts for Change (2013) textbook. These examples were chosen because they were based on real data (Hamlin, Wynn, & Bloom, 2007; Internet Legal Research Group, n.d.), presented in textbooks that include simulation-based methods, and easily adaptable to statistical literacy tasks.

Lastly, the reviewers were provided with a test blueprint review form (see Appendix B3). The preliminary test blueprint was embedded within the form and included multiple questions. Reviewers were asked to rate how important each learning outcome was in determining how statistically literate a student is. A 4-point scale was used where 1 represented a learning outcome that was not essential for the assessment and 4 represented a learning outcome that was essential for the assessment. The remaining questions for the reviewers asked them to (a) provide additional comments on topics and learning outcomes that they felt were not clearly described, (b) list important topics and learning outcomes they felt were missing, and (c) provide additional suggestions.

Development of the first version of the test blueprint. Based on reviewers' feedback, modifications were made to create the BLIS Test Blueprint-1. The reviewers' feedback was compiled into one Microsoft Excel spreadsheet in order to examine feedback from all reviewers by learning outcome. For each learning outcome, the percentages of reviewers who rated the learning outcome a 1, 2, 3, and 4 were computed. These percentages were used to put the learning outcomes into four groups. The learning outcomes with similar ratings from the reviewers were grouped together. Almost all learning outcomes in the two groups with the highest ratings, Groups A and B, were kept in the test blueprint. Learning outcomes in group with the lowest ratings, Group D, were removed or modified. Learning outcomes in the remaining group, Group C, were examined in detail to determine if they should be removed or kept. Considerations included the emphasis of these outcomes in the four textbooks that were examined and written comments from the reviewers.

In addition to looking at the ratings from reviewers, comments from the reviewers were examined. Proposed additions to the topics and learning outcomes were evaluated by determining if the proposed topic or learning outcome aligned with the definition of statistical literacy that the assessment was designed to measure. Suggestions from two faculty members at the University of Minnesota, Dr. Joan Garfield and Dr. Michelle Everson, were also considered. After all changes were made, the BLIS Test Blueprint-1 was used to create the assessment.

Assessment Development

Multiple steps were taken to create the BLIS assessment using the test blueprint. The language of items was examined carefully. Items from existing instruments as well as items developed specifically for the BLIS assessment were used to create the preliminary version of the assessment. The following subsections describe each of the steps taken to create the different versions of the BLIS assessment.

Item writing considerations and item characteristics. The first step in creating the BLIS assessment was to decide what item format should be used: selected-response or constructed-response. Multiple arguments have been made for using constructed-response items rather than using selected-choice items. According to Downing (2006b) and Moreno, Martínez, and Muñiz (2006), selected-response items can be more difficult to write compared with constructed-response items. In addition, McGuire (1993) and Newble, Baxter, and Elmslie (1979) claimed that the amount of information gained from a selected-response item is not of much use and that constructed-response items give you much more information.

There are also concerns about using constructed-response items, and many arguments for using selected-choice items. The expense of scoring constructed-response items can be quite high and it takes time to score the responses (Downing, 2006a). Two potential problems with constructed-response items are that some participants' responses are vague, and written responses are difficult to evaluate objectively (Weathington et al., 2010). Rodriguez (2003) conducted a meta-analysis of studies comparing selected-response and constructed-response items and found evidence that if the stems for selected-response and constructed-response items were the same, the correlation between the scores was high. Objective tests with selected-response items have been the most widely used forms of assessment in standardized testing (Downing, 2006a). In addition, Downing (2006a) claimed that "selected-response items are the most appropriate item format for measuring cognitive achievement or ability" (p. 288). According to Downing, validity evidence for the selected-response item format is strong; varying ability levels can be assessed to reduce construct underrepresentation and subjectivity in scoring can be eliminated. For these reasons, the BLIS assessment will contain selected-response items.

In order to modify existing selected-response items and write new items, several item writing recommendations were consulted. First, item writing recommendations provided by the *Standards* (AERA, APA, & NCME, 1999) and Haladyna, Downing, and Rodriguez (2002) were considered. For example, words were avoided that have a different meaning or connotation for individuals of different ethnicities (AERA, APA, & NCME, 1999), and the central ideas were included in the stems instead of the choices (Haladyna et al., 2002). In addition, when creating selected-response items, constructed-response items were piloted first in order to choose appropriate distractors for the

selected-response options as recommended by Haladyna et al. (2002), Garfield and Franklin (2011), and Thorndike and Thorndike-Christ (2010).

Additional characteristics that are specific to assessments of statistical knowledge were considered when choosing and creating items. First, items included a real-world context as recommended by Gal (1998) and Garfield et al. (2005). Secondly, in order to create items to measure statistical literacy, the wording was carefully chosen to ensure the primary outcome being assessed is statistical literacy rather than other outcomes such as statistical reasoning or statistical thinking. Key words that can be used to assess statistical literacy, reasoning, and thinking were provided by Garfield et al. (2010) and were based on the key words mentioned by delMas (2002) as well as Garfield, et al. (2003). Key words to include when assessing statistical literacy are: *identify, describe, translate, interpret, read, and compute*. These words were used not only to help determine which items in existing assessments measure statistical literacy, but to create new items. More emphasis was placed on descriptions and interpretations and less emphasis was placed on computing because many courses have shifted focus away from computations to interpretations (Chance, 1997).

Development of the preliminary assessment. Using the item writing considerations and item characteristics just described, the preliminary statistical literacy assessment was created and reviewed. The assessment was a combination of items from existing instruments, revised items from existing instruments, and newly created items. See Appendix C1 for a copy of the preliminary assessment.

Nineteen existing items measuring statistical literacy were chosen from the CAOS test (delMas et al., 2007), ARTIST Topic Scale tests (Garfield et al., 2002), ARTIST item

database (Garfield et al., 2002), and an early version of the GOALS assessment (Garfield et al., 2012). These particular assessments were examined because the assessments included items measuring statistical literacy, contained items with real-world contexts, and were created for introductory statistics students. The CAOS test was chosen specifically because of its relatively high reliability. The ARTIST Topic Scale tests were chosen because a majority of the items measure statistical literacy. The ARTIST item database was examined because it has items already put into categories of statistical literacy, reasoning, and thinking. Lastly, an early version of the GOALS assessment was chosen because it was created for students in an introductory statistics course that includes simulation-based methods. Items were chosen that match the topics and learning outcomes listed in the test blueprint. The selected items were used as they were written or modified to match the learning outcomes specified in the test blueprint. For three of the items that included a graph, the graphs were re-constructed in R version 3.0.1 (R Core Team, 2013) using the *plotrix* package (Lemon, 2006) so they would be similar in format. See Table 7 for a summary of the items chosen from existing instruments.

Table 7

BLIS Assessment Items Chosen from the CAOS Test, Artist Topic Scale Tests, Artist Item Database, and the GOALS Assessment Matched with Their Learning Outcomes and Sources

Item	Learning outcome	Source
4	Ability to determine what type of study was conducted	ARTIST Topic Scale Test: Data Collection - Item 7 ^a
5	Ability to determine if a variable is quantitative or categorical	ARTIST Topic Scale Test: Data Collection - Item 2
6	Ability to determine if a variable is an explanatory variable or a response variable	ARTIST Topic Scale Test: Data Collection - Item 3
7	Understanding of the difference between a statistic and parameter	ARTIST Item Database - Item Q0618
9	Ability to describe and interpret a dotplot	ARTIST Topic Scale Test: Data Representation - Item 12 ^a
10	Ability to describe and interpret the overall distribution of a variable	CAOS - Item 1
12	Ability to interpret a probability in the context of the data	ARTIST Topic Scale Test: Probability - Item 9
14	Understand how a mean is affected by skewness or outliers	ARTIST Topic Scale Test: Sampling Variability - Item 7
15	Ability to interpret a standard deviation in the context of the data	ARTIST Topic Scale Test: Measures of Spread - Item 2
16	Understanding of the properties of standard deviation	ARTIST Topic Scale Test: Measures of Spread - Item 5
17	Understanding of what an empirical sampling distribution represents	ARTIST Topic Scale Test: Sampling Variability - Item 1 ^a
21	Understanding that a confidence interval for a proportion is centered at the sample statistic	ARTIST Item Database - Item Q0943 ^a
22	Understanding of how the confidence level affects the width of a confidence interval	GOALS - Item 14 ^a
27	Understanding of the logic of a hypothesis test	CAOS - Item 40
31	Ability to determine statistical significance based on a p-value	CAOS - Item 19
34	Understanding that only an experimental design with random assignment can support causal inference	ARTIST Topic Scale Test: Data Collection - Item 4
35	Understanding of the factors that allow a sample of data to be generalized to the population	CAOS - Item 38
36	Ability to match a scatterplot to a verbal description of a bivariate relationship	GOALS - Item 7
37	Ability to use a least-squares regression equation to make a prediction	ARTIST Topic Scale Test: Bivariate Data, Quantitative - Item 13

^aItem was modified from its original version

For the 18 learning outcomes that did not have existing items available, new items were created. Textbooks (Gould & Ryan, 2013; Lock et al., 2013) and journal articles (e.g., Utts, 2003) were examined to find real-world contexts that could be used to create items. Contexts were chosen purposively to be interesting to students and not biased towards certain populations. For example, contexts were not chosen that were related to sports since this might result in some students being more interested in the item than other students. Multiple contexts were chosen from the Pew Research Center (n.d.) surveys, which were used in the Lock et al. (2013) textbook and the Gould and Ryan (2013) textbook. Media articles from websites, which were tied back to the journal article in which the study was originally published in, and other online sources, were used. For some contexts, items were written to create testlets to reduce the amount of reading students had to do in order to answer the questions. A testlet is a set of items that included a common stem (Downing, 2006a). See Table 8 for a list of item sources that were chosen by item.

Table 8

Ne Newly Written BLIS Assessment Items Matched with Their Learning Outcomes and Sources

Item	Learning outcome	Source
1	Understanding of the difference between a sample and population	Pew Research Center (2013)
2	Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term	N/A
3	Understanding that statistics computed from random samples tend to be centered at the parameter	N/A
8	Understanding that statistics vary from sample to sample	Gould and Ryan (2013)
11	Understanding the importance of creating graphs prior to analyzing data	Beishe et al. (2004) from Gould and Ryan (2013)
13	Ability to interpret a mean in the context of the data	N/A
18	Understanding that an empirical sampling distribution shows how sample statistics tend to vary	Pew Research Center (2011)
19	Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter	Pew Research Center (2011)
20	Understanding that a confidence interval provides plausible values of the population parameter	Duggan (2013)
23	Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis	Mednick, Cai, Kanady, & Drummond, (2008) from Gould and Ryan (2013)
24	Understanding of how sample size affects the standard error	Mednick et al. (2008) from Gould and Ryan (2013)
25	Understanding that a randomization distribution tends to be centered at the hypothesized null value	Steinberg, Levine, Askew, Foley, and Bennett (2013)
26	Ability to estimate a p-value using a randomization distribution	Steinberg et al. (2013)
28	Understanding of the purpose of a hypothesis test	Chervin et al. (2013) from Science World Report (n.d.)
29	Understanding that every model is based on assumptions which limit our scope of inferences	delMas (2013)
30	Ability to determine a null and alternative hypothesis statement based on a research question	Amherst H. Wilder Foundation (n.d.)
32	Understanding that errors can occur in hypothesis testing	Utts (2003)
33	Understanding of how a significance level is used to make decisions	Sonoda et al. (2011) from Lock et al. (2013)

Almost all of the new items were written in a constructed-response format in order to use student responses in the pilot of the assessment to discover plausible incorrect answers to make meaningful distractors. One new item that was in a selected-response format on the preliminary assessment was not in a constructed-response format because the purpose of the item was to compare sequences of numbers and choose which was most plausible for a sampling distribution. For the remaining new items, students were asked to explain their answers so that incorrect answers could be understood and judged in order to determine if they should be included as selected-response options in a later version of the assessment.

Expert review of the preliminary assessment. After the preliminary assessment was created, the same six reviewers who examined the test blueprint were asked to review the assessment. See Appendix D1 for the invitation that was sent to the reviewers, which was adapted from the invitation to reviewers for the test blueprint. For reference, reviewers were provided with a summary of the changes to the test blueprint (see Appendix D2). They were also provided with a review form for the preliminary assessment (see Appendix D3). Similar to the review form for the test blueprint, the assessment items were embedded in the review form. In addition, for each item, the learning outcome was stated. For the existing items, the reviewers were told which assessment the item came from.

For each item, reviewers were asked how much they agreed or disagreed with the following statement: “The assessment item measures the specified learning outcome.” They rated their agreement on a 4-point scale: 1=strongly disagree, 2=disagree, 3=agree, and 4=strongly agree. Below each item, reviewers were provided with space to provide

comments. At the end of the review form, reviewers had the option to provide any additional suggestions for improving the assessment.

Development of the first version of the assessment. Similar to what was done with the test blueprint, modifications were made to reflect the suggestions of the reviewers, and this resulted in the BLIS-1 assessment (see Appendix C2). For two of the items taken from existing instruments that included a graph, the graphs were reconstructed in R using the *plotrix* and *ggplot2* packages (Lemon, 2006; Wickham, 2009) so they would be similar in format. The item ratings were again summarized in a Microsoft Excel sheet and grouped using the same criteria as the ratings for the test blueprint. For a handful of items, one reviewer gave two ratings instead of one. The ratings, however, still fit in with the grouping specifications and as a result, there were no problems putting those items into groups. All comments from reviewers were compiled into a Microsoft Word document. Reviewers had been asked to comment on the items, but one reviewer also suggested a minor wording change to one of the learning outcomes. The suggested change was made to the test blueprint, which was renamed as the BLIS Test Blueprint-2.

Cognitive interviews with students. In order to make the second version of the assessment, the following steps were taken. Interviews were conducted with six students; four were conducted face-to-face and the other two were conducted using Skype with video. All interviews were audio-taped. Students were asked to talk about what they are doing and thinking while taking the BLIS-1 assessment. Student responses were examined to determine if there were any problematic items. The following subsections describe the details of each step.

Participants. Students were recruited to participate in cognitive interviews in one of two ways: in-class invitations and emails distributed to students by their instructors. For the in-class invitation, a script was prepared (see Appendix E1) and a sign-up sheet (see Appendix E2) was passed around for students who were interested in participating. For the students contacted by email, the invitations, which were modified from an invitation written by Park (2012), requested students to email the dates they would be available to participate. Two different invitation letters were sent out, one requesting students to participate in-person and the other requesting non-Minnesota students to participate using Skype (see Appendices E3 and E4, respectively). The invitations that were emailed to students were sent by their instructors. The instructors were either personally contacted or solicited in an email sent to an electronic mailing list for a group of statistics educators in Minnesota. See Appendix E5 for the letter that was sent to request instructors to email their students the invitation to participate in the cognitive interviews.

Cognitive interviews were conducted with six students from four courses. Two graduate students at the University of Minnesota who took a graduate-level introductory statistics course, taught with the *Statistics: Unlocking the Power of Data* (Lock et al., 2013) textbook, agreed to participate. Two undergraduate students at the University of Minnesota who completed the CATALST course (Garfield, et al., 2012), which used the *Statistical Thinking: A Simulation Approach to Modeling Uncertainty* (Catalysts for Change, 2013) textbook, volunteered to participate. In addition, two undergraduate students from California Polytechnic State University agreed to be interviewed: one took a course that used the *Introduction to Statistical Investigations* (Tintle et al., 2013)

textbook and the other student took a course that used the *Investigating Statistical Concepts, Applications, and Methods* (Chance and Rossman, 2006) textbook. Each student received a \$20 gift card to Amazon.com as an incentive to participate. The interviews were conducted three to ten weeks after the students completed their introductory statistics course.

Interview procedures. Prior to conducting cognitive interviews with students, an interview protocol was developed. The protocol was a modified version of an example interview protocol provided by Willis (2004; see Appendix E6). After the protocol was modified, it was reviewed by a graduate student studying statistics education who had recently conducted cognitive interviews for another assessment development project. The protocol informed students that they should think aloud as they took the assessment by saying everything they were thinking about as they answered the questions. Students were walked through two example questions. The protocol also specified that while the student took the assessment, the interviewer would not interrupt the student unless to ask for clarifications on the student's answers. Furthermore, the interviewer was not to answer any content related questions. Willis (2004) claimed the benefit of conducting an interview in this manner is that there is little, if any, interviewer-imposed bias.

The four students from the University of Minnesota were individually interviewed face-to-face in an office with no other individuals present. At the beginning of the interview, the students were provided with a consent form that was modified from Park's (2012) consent form (see Appendix E7) and asked to read and sign it. After students signed the consent form, they were provided with a copy of the consent form for their own reference. The interviewer then followed instructions in the interview protocol.

The two interviews that were conducted via Skype were done similarly to the face-to-face interviews with one main difference. A couple of days before the interview, students were emailed an electronic copy of the consent form and were asked to provide an electronic signature prior to the interview.

Development of the second version of the assessment. Student responses from the cognitive interviews were used to make changes to the BLIS assessment. Changes were made if a student appeared to be misguided by the wording of the item or if a student had to re-read an item multiple times. In addition, any minor mistakes in language were fixed.

After making changes based on the results from the student interviews, multiple meetings took place with one of the six assessment reviewers as well as two faculty members at the University of Minnesota, Dr. Joan Garfield and Dr. Michelle Everson. Modifications made to the assessment were discussed, in addition to particular concerns about items that appeared to lead students in the wrong direction. Collaboratively, decisions were made on whether any additional changes were needed. Lastly, the assessment, BLIS-2, was copied onto an online survey platform, Qualtrics, to be administered to students in the pilot test. See Appendix C3 for a copy of the BLIS-2 assessment.

Pilot administration. The next step was to run a pilot test, which was administered by three statistics instructors in Minnesota. Personal communication was used to recruit an instructor from the University of Minnesota to administer the assessment to her introductory statistics students in addition to the students enrolled in an introductory statistics course taught by the author. According to the *Standards* (AERA,

APA, & NCME, 1999), it is important that the sample used in a pilot should be as representative of the population as possible. Therefore, multiple attempts were made to encourage other instructors to participate in the pilot. In a first attempt to recruit instructors from other institutions, a request was made at a monthly meeting of statistics educators in Minnesota. More attempts were made by emailing specific instructors in Minnesota an invitation. The invitation letter was based loosely on the invitation that was sent to the reviewers (see Appendix F1). One instructor that responded with interest was asked to participate in the pilot test.

The instructors who agreed to administer the assessment to their students were provided with instructions on how to administer the assessment to their students (see Appendix F2), which included a link to the assessment and a unique code that students would use in the assessment to identify which class they were in. In addition, the instructions included a link to a survey on Qualtrics, which was meant to be filled out by the instructors (see Appendix F3). The survey was designed to gather characteristics of the courses students were enrolled in, including the state their institution is in, the type of institution, the mathematics prerequisites for the course, whether the assessment was administered in or out of class, and if students received any credit for taking the assessment. These questions were based on a similar set of questions that was given to instructors who administered the CAOS test (delMas et al., 2007). The purpose of administering the survey with the pilot test was to see if there were any problematic questions before administering the survey with the field test.

There were 76 students enrolled in three introductory statistics courses who took the assessment. The introductory statistics courses included a graduate-level course

taught with the *Statistics: Unlocking the Power of Data* (Lock et al., 2013) textbook at the University of Minnesota, an undergraduate-level course taught with the CATALST curriculum (Garfield, et al., 2012) at the University of Minnesota, and an undergraduate-level course taught with *Statistics* (McClave & Sincich, 2007) at Winona State University. Students at the University of Minnesota completed the assessment outside of class and were offered extra credit to complete the assessment. Students at Winona State University completed the assessment in class and were given credit towards their grade for completing the assessment.

Students completed the assessment using the online survey platform, Qualtrics. A consent form was included in the online assessment on the first page (see Appendix F4) which was based on a consent form used by Park (2012). After seeing the consent form, students were prompted to enter their name and a code that was given to their instructor to identify their class. A short set of instructions (see Appendix F4) was then provided and students were then able to complete the assessment. Lastly, students were asked to provide some demographic information (see Appendix F5), which was not required. Students were asked what their gender was, what age category they belonged to (where categories were in five year increments), their class level, whether they were an international student or foreign national student, and their racial identification. The categories for the question about racial identification were based on a question from the American Community Survey (U.S. Census Bureau, 2014). The purpose piloting the demographic questions was to see if any were problematic before doing the field test. Demographic information was not used in analyzing students' performance on the pilot test.

Development of the third version of the assessment. The students' results from the pilot test were examined to make changes to the assessment. First, the students' responses were downloaded into a Microsoft Excel document and the data was cleaned. The data cleaning included removing data from non-consenting students, deleting duplicate responses from the same student, and deleting responses for students' who did not answer any of the items (i.e., answered the consent question but no content questions).

The student responses for the selected-response items were analyzed using R. The total number of students who chose each response, as well as the percentage of students who chose each response, were computed. The percentage of students who chose each selected-response option was computed for each course separately as well as overall. The purpose of computing these percentages was to determine if all of the selected-response options were needed. If no students chose a particular option, it was considered for deletion. In addition, if the percentage of students who chose a particular distractor was higher than the percentage of students who chose the correct answer for an item, the item was examined in more detail.

For the constructed-response items, student responses were examined to convert constructed-response items to selected-response for the third version of the BLIS assessment (BLIS-3). In order to create the selected-response options, students' responses were grouped by similar responses. The incorrect answers that were submitted most often and appeared to measure misconceptions related to the learning outcomes were included as distractors. See Appendix C4 for a copy of the BLIS-3 assessment.

Field test administration. Instructors who taught an introductory statistics course that included simulation-based methods at the college-level, which included Advanced Placement Statistics courses, were recruited to administer the BLIS-3 assessment to their students. The invitation letter contained a description of the information the participating instructors would gain, such as information about what knowledge their students have and how they compare with students at other institutions, in order to provide an incentive for them to participate (see Appendix G1). Using the invitation letter, statistics instructors were recruited by personal emails, through the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) website (<http://www.causeweb.org>), and through the Isolated Statisticians electronic mailing list (isoStat; <http://www2.lawrence.edu/fast/jordanj/isostat.html>). The administrators for the CAUSE electronic mailing list requested modifications to the invitation letter and therefore a second version of the letter was created (see Appendix G2).

Instructors who were interested in using the assessment were sent a letter which provided them with more information about the assessment, including the format of the assessment and how it would be administered (see Appendix G3). If the instructors were still interested, they were provided with a set of instructions for administering the assessment to their students, which included a link to the assessment and a code that students would enter to identify their class. The set of instructions was based on a similar set of instructions written by Park (2012). See Appendix G4 for the final set of instructions.

The instructors who agreed to administer the assessment to their students were also asked to fill out a short survey (see Appendix G5). The survey was designed to

gather characteristics of the courses students were enrolled in, including the country their institution is in, the type of institution, the mathematics prerequisites for the course, whether the assessment was administered in or out of class, if students received any credit for taking the assessment, if the assessment was administered in the beginning, middle, or end of their course, and what type of simulation methods were used in the course (see Table 9). All, except for two, courses were in the United States and most incorporated simulations in the curricula. Most students took the assessment out of class and received extra credit for completing the assessment. In addition, the survey included a question asking them to rate how valuable they believe the BLIS assessment is for statistics educators. The value to instructors and value to educational researchers were rated separately on a 4-point scale from “Not at all valuable” to “Very valuable.”

Table 9

Course Characteristics as Reported by Instructors

Characteristic	Number of courses (N=34)		Number of students (N=930)	
	<i>n</i>	%	<i>n</i>	%
Country				
United States	32	94.1	903	97.1
Canada	1	2.9	26	2.8
Spain	1	2.9	1	0.1
Type of institution ^a				
High school	3	9.1	114	12.4
2-year/technical college	6	18.2	126	13.7
4-year college	12	36.4	294	31.9
University	12	36.4	387	42.0
Pre-requisites ^{bc}				
No mathematics	4	12.9	128	13.8
High school algebra	16	51.6	374	40.2
College algebra	10	32.3	377	40.5
Calculus	3	9.7	36	3.9
Setting ^d				
In class	5	16.7	182	21.8
Out of class	25	83.3	652	78.2
Credit ^b				
Assignment	5	16.1	128	14.7
Extra credit	23	74.2	682	78.5
No credit	3	9.7	59	6.8
Time of semester ^b				
Beginning	0	0.0	0	0.0
Middle	6	19.4	177	20.4
End	25	80.6	689	79.6
Simulation methods ^{bc}				
Bootstrapping	10	32.3	348	37.4
Randomization tests	14	45.2	466	50.1
Probability simulations	24	77.4	716	77
Other simulations	23	74.2	674	72.5
None of the above	4	12.9	77	8.3

Note. Ten students completed the assessment but did not provide their instructor code so their results are not included in this table.

^aOne instructor did not respond, ^bThree instructors did not respond, ^cPercentages do not add up to 100 because instructors could choose more than one option, ^dFour instructors did not respond.

Students who completed the BLIS assessment were provided with the same consent form that was used in the pilot administrations. Students were asked to provide some demographic information, which was at the end of the assessment and optional. The demographic questions were the same questions that were included in the pilot. The question asking students which class they belonged to did not include a *high school* option, so the students' results were linked to their institution and responses for students in high school were changed. See Table 10 for the student demographics from the field test. A majority of students (87.9%) were under the age of 25 and most students (87.4%) were in college.

Table 10

Demographic Characteristics of Students Who Took BLIS-3 in the Field Test

Characteristic	<i>n</i>	%
Gender		
Female	533	58.3
Male	382	41.7
Age		
19-	364	39.9
20-24	438	48.0
25-29	55	6.0
30-34	21	2.3
35-39	15	1.6
40-44	9	1.0
45-49	5	0.5
50-54	3	0.3
55+	2	0.2
Class		
High school	114	12.6
Freshman/first year	190	21.1
Sophomore	291	32.3
Junior	166	18.4
Senior	90	10.0
Graduate student	31	3.4
Other	20	2.2
International or foreign national student		
Yes	44	4.8
No	868	95.2
Race		
White	710	82.8
Black or African American	24	2.8
American Indian or Alaska Native	7	0.8
Asian	66	7.7
Pacific Islander	6	0.7
Other	44	5.1

Analysis of field test data. In this sub-section, the analyses that were conducted to examine the reliability, validity, and value of the BLIS assessment are described.

Analyses based on Classical Test Theory (CTT) and Item Response Theory (IRT) were

used to check for reliability. The analyses that were performed in order to check for validity were based on IRT. A measurement expert at the University of Minnesota was consulted with to ensure appropriate analyses were conducted. Instructors' feedback from the field test was used to determine the value of the assessment.

Descriptive statistics for the BLIS-3 assessment. Descriptive statistics were computed for the BLIS-3 assessment. First, the percentages of students who chose each selected-response option were computed using the *gmodels* package in R (Warnes, 2013). The percentages were examined to see if any of the items had a higher percentage of students who selected a particular distractor than the correct option. Then, total scores were computed and a plot of students' total scores was created using the *plotrix* package in R (Lemon, 2006).

Collecting further evidence of the reliability of the BLIS assessment. To check for score reliability, analyses based on CTT and IRT were used. In order to collect reliability evidence for the raw total scores of an assessment, analysis based on CTT was conducted (Hambleton, Swaminathan, & Rogers, 1991). Coefficient alpha, based on CTT, was computed because it is a measure of internal consistency used when the items have varying difficulty levels and provides reliability evidence for the raw scores of the assessment (Thorndike & Thorndike-Christ, 2010). The *psych* package in R was used to compute the coefficient alpha statistic (Revelle, 2014).

Considering that the BLIS-3 assessment contained item pairs in testlets, coefficient alpha was computed using testlet scores. Each testlet contained a pair of items and was scored as a 0, 1, or 2, where a score of 0 indicated that both items in the testlet were incorrect and a score of 2 indicated that both items in the testlet were correct. If the

testlets are considered to be items, the number of items decreased from 36 to 32. Changing the number of items affects coefficient alpha (Sireci, Thissen, & Wainer, 1991) so the Spearman-Brown formula was used to predict what coefficient alpha would be for an assessment with 32 items and testlets, but with the same reliability characteristics as the assessment with 36 items. The predicted coefficient alpha, computed using the *CTT* package in R (Willse, 2014), was compared to the actual coefficient alpha for the assessment data incorporating testlet scores to examine if local item dependence existed for items in testlets.

To check for reliability at different ability levels and at the item level, analysis based on IRT was conducted. All analysis based on IRT was conducted using the *ltm* package in R (Rizopoulos, 2006). The test information function and standard error of measurement summarized how well the assessment measured the construct at different ability levels (AERA, APA, & NCME, 1999). In order to examine individual item reliability, item information curves were created.

The required assumptions needed to conduct analyses based on IRT were checked as outlined by Raykov and Marcoulides (2010). The first assumption examined was that there was one dimension or one construct (i.e., statistical literacy). A single-factor CFA was conducted using Mplus (Muthén & Muthén, 2010) to provide evidence that the score results from the assessment were unidimensional. The weighted least squares with mean and variance adjustment (wlsmv) was used as the estimation method, which was recommended by Muthén, DuToit, and Spisic (1997) for data with categorical responses. The following results from the single-factor CFA were examined: a scree plot of eigenvalues, factor loadings, and model fit indices.

The second assumption checked was local independence, which means that each item is independent of any other item and uncorrelated after controlling for other factors. Local dependence can occur when there is multidimensionality that is unaccounted for or if there are other kinds of dependency such as having items in a testlet (Downing, 2006a). The BLIS assessment has five testlets, each with two items. Therefore, it was important to investigate if there were any local dependencies. In order to check the local independence assumption, the tetrachoric correlations residual were examined. Further, the single-factor CFA was rerun using testlet scores to examine if unidimensionality was still met.

Assuming the unidimensionality assumption was met, three IRT models were fit to the data: the Rasch model (Bond & Fox, 2007), 2 parameter logistic (2PL) model (Hambleton et al., 1991), and partial credit (PC) model (Masters, 1982). All models have a person ability parameter, θ , and a difficulty parameter, b . A discrimination parameter, a , is added to the 2PL model. The PC model incorporates testlet scores in the analysis while the Rasch model and 2PL model do not.

For the Rasch and 2PL models, the probability of a randomly chosen person with a particular ability level, θ , getting item i correct for each of the models is given by

$$\text{Rasch: } P_i(\theta) = \frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}} \quad i = 1, 2, \dots, n$$

$$\text{2PL: } P_i(\theta) = \frac{e^{D a_i (\theta-b_i)}}{1+e^{D a_i (\theta-b_i)}} \quad i = 1, 2, \dots, n$$

where i is the item number, D is a scaling factor assuming the value of 1.7, and n is the number of items incorporated in the model. For the PC model, the probability of a randomly chosen person with a particular ability level, θ , scoring a 0, 1, or 2 on a particular item or testlet i for each of the models is given by

$$\text{PC: } P_{ik}(\theta) = \frac{\exp \sum_{j=0}^k (\theta - b_{ik})}{\sum_{i=0}^{m-1} (\theta - b_{ik})} \quad i = 1, 2, \dots, n$$

where i is the item or testlet number, k is the category ($k = 0, 1, \dots, m$), and n is the number of items and testlets incorporated in the model. The Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Log Likelihood were used to choose the best fitting model. The best fitting model was used to conduct the remaining analyses based on IRT.

Collecting evidence of the validity of the BLIS assessment. In order to determine the extent to which the interpretations of the BLIS assessment as a measure of statistical literacy have validity, multiple methods were used to create one holistic argument. First, the content of the assessment was examined using qualitative methods. The expert reviewers examined and gave feedback on the test blueprint and assessment, which, according to Thorndike and Thorndike-Christ (2010) and Weathington et al. (2010), provides evidence of validity. As mentioned previously, statistics instructors and researchers were chosen as reviewers. Having reviewers with a variety of backgrounds provided more evidence of validity (Thorndike & Thorndike-Christ, 2010). Feedback provided by the reviewers informed whether the items measure the topics presented in the test blueprint and whether the assessment as a whole measures statistical literacy. The cognitive interviews with students indicated whether the BLIS-1 assessment was measuring the intended content providing evidence of validity.

Quantitative methods that were used to measure validity were CFA and analyses based on IRT. CFA has been used to determine whether or not an assessment measures one underlying construct (Thorndike & Thorndike-Christ, 2010). Therefore, CFA was used in this study to provide evidence that the assessment is measuring the one intended

construct, statistical literacy. Next, analyses based on IRT were conducted to determine the extent to which an assessment has internal validity (Thorndike & Thorndike-Christ, 2010). Item difficulties were calculated to display whether or not the assessment included items of varying difficulty levels.

Collecting evidence of the value of the BLIS assessment. Instructors who administered the BLIS-3 assessment to their students in the field test provided evidence of the value of the assessment. As mentioned previously, in the survey that was given to the instructors, they were asked to rate how valuable they believe the BLIS assessment is for statistics educators. The value to instructors and value to educational researchers were rated separately on a 4-point scale from “Not at all valuable” to “Very valuable”. The total number of instructors who gave each rating were computed as well as the percentage of instructors who gave each rating.

Chapter Summary

This chapter outlined the development process of the BLIS test blueprint and BLIS assessment. The collection of reliability and validity evidence were described as well as how the value of the assessment will be assessed. The next chapter will show the results of the study.

Chapter 4

Results

This chapter describes the results collected during the development of the BLIS assessment. Feedback from the expert reviewers is described, and the changes made to the BLIS test blueprint and assessment based on the feedback are reported. Student responses from the cognitive interviews and small-scale pilot test are then presented along with the changes that were made to create the BLIS-2 and BLIS-3 assessments. The results from the analyses of the data collected during the large-scale field test are then presented.

Results from Expert Review of the Test Blueprint

As mentioned previously, the preliminary test blueprint was created based on a review of introductory statistics textbooks (Gould & Ryan, 2013; Catalysts for Change, 2013; Lock et al., 2012; Tintle et al., 2013) that incorporate simulation-based methods. A total of 26 topics were included in the preliminary test blueprint. Each topic had at least one learning outcome, and, in all, 54 learning outcomes were devised. The learning outcomes were based on the review of textbooks. If there were multiple learning outcomes that were similar, the outcomes incorporated in more textbooks were included. For example, consider these two learning outcomes: the ability to describe and interpret a dotplot, and the ability to describe and interpret a histogram. These two learning outcomes were very similar, and because dotplots were emphasized in more textbooks than histograms were emphasized, the learning outcome about dotplots was included instead of the learning outcome about histograms. See Table A1 in Appendix A for the preliminary test blueprint.

The preliminary test blueprint was sent out to be reviewed by the expert reviewers described in the previous chapter. Using the comments from the expert reviewers, many changes were made to create the BLIS Test Blueprint-1 (see Table A2 in Appendix A). Each learning outcome was rated on how essential the learning outcome was for the assessment. A majority of the learning outcomes appeared to fall naturally into four groups. The highest rated group, Group A, contained 30 learning outcomes with at least four 4's, or at least three 4's with no 1's. Recall that a rating of 4 represented a learning outcome that the reviewers believed essential to be included on the assessment and a 1 represented a learning outcome that was not essential to be included. Group B, the second highest rated group, included nine learning outcomes. Some of Group B's learning outcomes included at least three 4's with 1's, and for the other learning outcomes, over half of the ratings were 3's and 4's, excluding learning outcomes in Group A. The third group, Group C, included nine learning outcomes that were not included in the other groups. Group D included seven items where over half of the ratings were 1's, 2's, or no rating. Table H1 in Appendix H includes the ratings from the six expert reviewers and the group that each learning outcome was put in.

Reviewers' comments were examined to better understand particular ratings. See Table H2 in Appendix H for the reviewers comments and how each comment was addressed. The first thing that stood out in the ratings was that 10 learning outcomes related to bootstrap distributions, and bootstrap intervals had much lower ratings compared to randomization distributions and randomization tests. Multiple reviewers said that not all introductory statistics courses that use simulation-based methods include bootstrap intervals. For example, the Tintle et al. (2013) textbook includes randomization

tests, but not bootstrap intervals. Therefore, the four bootstrap learning outcomes in Group D were removed. One of the learning outcomes in Group C, understanding of the purpose of a bootstrap interval, was removed because it overlapped with a different learning outcome in Group B, understanding that a bootstrap interval provides plausible values of the population parameter. Another learning outcome in Group C, which involved describing a model for a bootstrap interval, was removed because one reviewer mentioned that *modeling* was too curriculum specific and another reviewer did not provide a rating because she did not understand what modeling represented in a bootstrap interval context. For the remaining four bootstrap learning outcomes, they were changed to refer to sampling distributions or confidence intervals that were not specific to a particular distribution.

Other learning outcomes were removed from the test blueprint for a variety of reasons (see Table H2 in Appendix H). All learning outcomes in Group D were deleted. Five learning outcomes from Group A were deleted. One learning outcome in Group A, the ability to interpret a margin of error, was removed because it was decided this learning outcome was measuring statistical reasoning rather than measuring statistical literacy. Four learning outcomes from Group A were deleted because they overlapped in content with other learning outcomes. For example, there were two learning outcomes about random assignment and cause-and-effect statements. For two of the overlapping pairs of learning outcomes, the one with lower ratings was deleted. For the other three, a comparison of the clarity in the language of the learning outcome and language used in the textbooks was used to determine which learning outcome in the overlapping pairs was

retained. One learning outcome in Group C was deleted because it also overlapped with another learning outcome in Group B.

Two learning outcomes in Group B, the ability to create an appropriate graph to display quantitative data and the ability to create a randomization distribution to test the difference between two groups, were deleted since students would not be able to create a graph or a distribution in a selected-choice type of item. Also, while taking the assessment, students will not necessarily have access to software that would permit the creation of a randomization distribution.

After it was decided that 33 learning outcomes would be retained, feedback provided by the expert reviewers was used to add four learning outcomes to the test blueprint. Suggested new learning outcomes were examined to see if they aligned with the definition of statistical literacy used to develop the assessment. Those learning outcomes that did align were added if they did not overlap with other learning outcomes already included in the test blueprint. Learning outcomes were added about model assumptions, errors in hypothesis testing, scatterplots, and making predictions using a least-squares regression equation.

After the learning outcomes were finalized, the topics were re-examined. One new topic was added, regression and correlation, to include the two new learning outcomes written about scatterplots and making predictions using a least-squares regression equation. In the preliminary test blueprint, many topics were similar. For the next version of the test blueprint, similar topics were combined. For example, the preliminary test blueprint included the following topics: samples and populations, randomness, random samples, observational studies and experiments, variables, and

statistics and parameters. Those topics were combined into one topic, data production. After the topics were finalized, Phase 1 of the assessment development was finished; the BLIS Test Blueprint-1 was complete (see Table A2 in Appendix A).

Results from Expert Review of the Assessment

Phase 2 of the assessment development involved creating the preliminary version of the assessment with 37 items, as described in Chapter 3, and conducting an expert review of the assessment. The same six reviewers of the test blueprint reviewed the preliminary assessment. The process used to make changes to the assessment based on the reviews was similar to the process used when making changes to the test blueprint. The same rules were used to categorize items into four groups, A through D. One reviewer checked two ratings instead of one rating for seven items. For example, the reviewer selected ratings of 2 and 3 for Item 10 instead of choosing one rating. These seven items still fit into the four groups. Items in all groups had changes, but items in Group C had the most changes. Only one item was in Group D, which was replaced with a new item. The following paragraphs describe these changes in more detail. Table I1 in Appendix I includes a summary of the ratings from the six expert reviewers and the group that each item was put in.

Many changes were made to the assessment items based on the reviewers' feedback. Comments from reviewers are included in See Table I2 in Appendix I. Reviewers suggested changes for almost all of the items. Changes that were made included grammatical changes, contextual changes, and statistical content changes. An example of a contextual change was one involving average weights of hamburger pizzas.

One reviewer thought that a *hamburger* pizza was “odd” and another reviewer claimed to have never heard of a hamburger pizza. The item was changed to refer to sausage pizza.

Multiple items had statistical content changes. One item, Item 14, provided students with a histogram and asked them to select a range of values that was most likely to include the sample mean. One reviewer noted that students often mistakenly believe that the mean should be close to the mode. Therefore, a selected-response option was created that had a range capturing the mode. Another item, Item 20, had two changes. First, the item referred to a bootstrap interval when it should have referred to a confidence interval. Second, the population was defined more specifically as including only American adults. Item 20 had other suggestions from reviewers that were not used because the item had high ratings; the item was in rating Group A. Item 34 was changed because it referred to types of studies that many introductory statistics courses do not cover (e.g., time series study). Also, similar to Item 20, Item 34 had an additional comment that was not used to make a change because the ratings were high for the other reviewers; Item 34 was in rating Group A.

As mentioned earlier, there was only one item, Item 12, in Group D, which was the group with the lowest ratings (see Table 11). The item was meant to measure the following learning outcome: ability to interpret a percent in the context of the data. Four reviewers provided comments about why they did not like the item such as “This is less about stat and more about meteorology” and “Even meteorologists disagree on what such a ‘probability’ means and some think of it as more subjective than objective.” A new item was written to replace the meteorology item and the learning outcome changed to the ability to interpret a probability in the context of the data (see Table 11).

Table 11

New Item for the BLIS-1 Assessment Created to Replace the Item with Low Ratings from Reviewers for the Preliminary Assessment

Old item in the preliminary assessment

Learning Outcome: Ability to interpret a percent in the context of the data

Item was taken from the ARTIST Topic Scale Test – Probability, Item 9

Item 12: The local Meteorological claims that there is a 70% probability of rain tomorrow. Provide the best interpretation of this statement.

- a. Approximately 70% of the city will receive rain within the next 24 hours.
- b. Historical records show that it has rained on 70% of previous occasions with the same weather conditions.
- c. If we were to repeatedly monitor the weather tomorrow, 70% of the time it will be raining.
- d. Over the next ten days, it should rain on seven of them.

New item in the BLIS-1 assessment

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

New item based on a real-world context from National Cancer Institute. (n.d.)

Item 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. Explain what the statistic, .15, means in the context of this report from the National Cancer Institute.

The BLIS Test Blueprint-2 was created for two reasons. First, the learning outcome changed for Item 12 as described previously. Second, one of the reviewers was concerned about the following learning outcome: understanding that a confidence interval is centered at the sample statistic. He said that confidence intervals are not always centered at the sample statistic, so the learning outcome was changed to: understanding that a confidence interval for a proportion is centered at the sample statistic. See Table A3 in Appendix A for a copy of the BLIS Test Blueprint-2.

The changes that were made to the assessment based on the expert reviews were discussed with one of the reviewers at the University of Minnesota as well as with Dr. Michelle Everson. A few more minor changes were made which resulted in the BLIS-1 assessment. Appendix I1 includes all 37 items in the BLIS-1 assessment and how they were changed from the preliminary assessment.

Results from Student Cognitive Interviews

The BLIS-1 assessment was administered, in the form of a cognitive interview, to six students who had recently taken an introductory statistics course that included simulation-based methods. Students' responses were used to make additional changes to the assessment items in order to create the BLIS-2 assessment. A description of the students' responses and changes that were made to the assessment are described in this sub-section.

The students who participated in the interviews varied in statistical ability and background. Two students were female and the remaining four were male. Five of the students finished the assessment in an hour or less, and one student only finished the first 28 questions in the allotted hour. Table J1 in Appendix J includes a copy of the items that were changed and the student comments that led to these changes. Items that had only minor changes in wording are not included in the appendix. Items that had the most significant changes are described below.

Multiple students thought that Item 3 was too easy. Also, students were focusing on the means that were presented in the selected-response options. The item was changed so that the selected-response options included dotplots with 20 sample means instead of a

list of five sample means. The mean and standard deviation were no longer included in the item. See Table 12 for a copy of the item and how it was changed.

Table 12

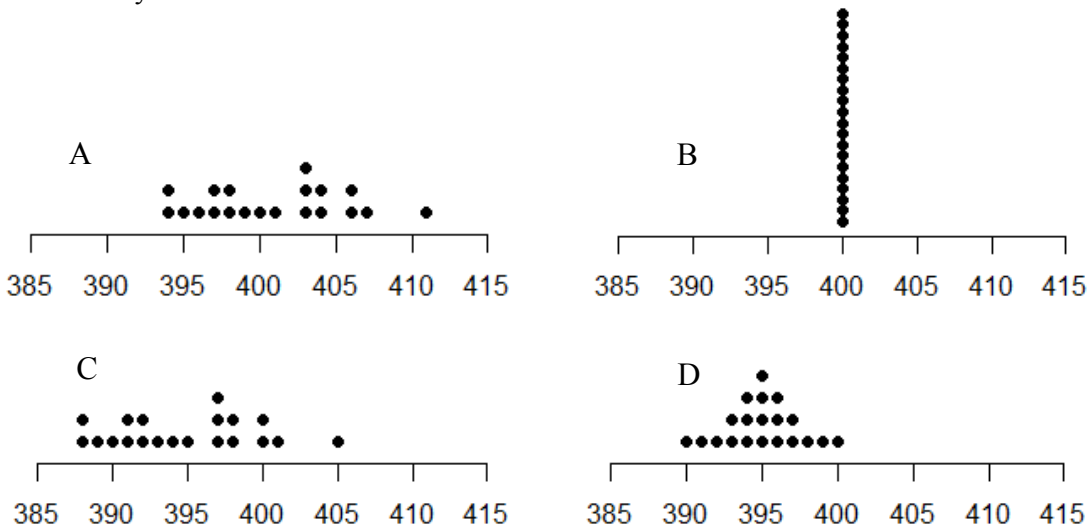
Changes Made to Item 3 After the Student Cognitive Interviews. Words that were Added are Underlined and Words that were Deleted are Crossed Out.

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

New item

Item 3: A manufacturer of frozen pizzas produces sausage pizzas, which have a true 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 ~~sequence~~ ~~below~~ is the most plausible for the average weights of 20 ~~the five~~ samples?

- a. ~~380, 385, 413, 424, 437 (mean = 407.8, sd = 24.67)~~
- b. ~~336, 362, 377, 387, 400 (mean = 372.4, sd = 24.64)~~
- e. ~~396, 400, 426, 445, 449 (mean = 423.2, sd = 24.63)~~
- d. ~~Any of the above.~~



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

Two items that were taken from ARTIST Topic Scale Tests had selected-response options that did not appear to be common student responses. For Item 9, students who were interviewed did not appear to use any statistical knowledge they would have learned in an introductory statistics class. Therefore, the selected-response items were deleted and the item was converted into a constructed-response item. The pilot administration of the assessment was then used to create better selected-response options. Item 17 was turned into a constructed-response item because it did not seem to contain good distractors. Many students thought that selected-response Option B in Item 17 was meant to be a *trick* option. See Table 13 for a copy of the items as they appeared in the BLIS-1 assessment along with students comments.

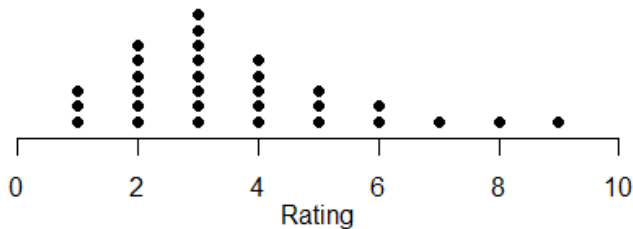
Table 13

Items 9 and 17 on the BLIS-1 Assessment that were Changed to be Constructed-Response Items for the BLIS-2 Assessment

Learning Outcome: Ability to describe and interpret a dotplot

Item was taken and modified from the ARTIST Topic Scale Test – Data Representation, Item 12

Item 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. The instructor examined the data for men and women separately. Below is the distribution of this variable for the 30 women in the class.



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

- A majority of women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
- The women in the class see themselves as having lower confidence of being able to succeed in statistics than the men in the class.
- If you remove the three women with the highest ratings, then the result will show an approximately normal distribution.

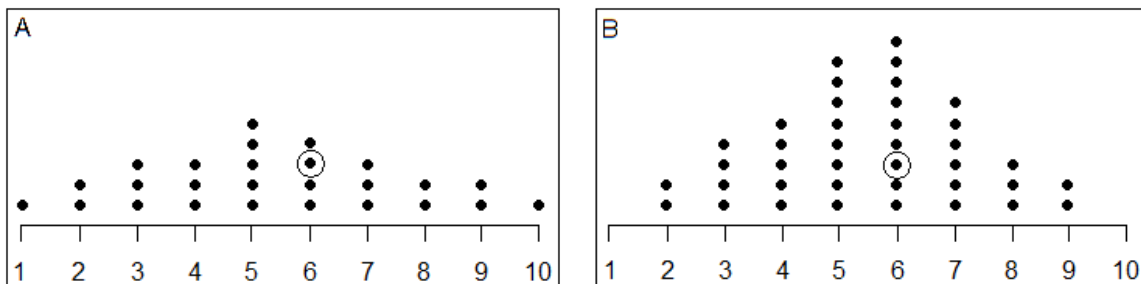
(continued)

Table 13 (continued)

Learning Outcome: Understanding of what an empirical sampling distribution represents

Item was taken and modified from the ARTIST Topic Scale Test – Sampling Variability, Item 1

Item 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.



- No, in both Figure A and Figure B, the circled dot represents one pebble that weights 6 grams.
 - Yes, Figure A has a larger range of values than Figure B.
 - 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.
-

Item 27 was taken from the CAOS assessment; however, another version of the item was created in a different assessment study (Ziegler, 2012). The original item from CAOS referred to electrical circuits as being “defective” as well as being “not good”. Ziegler (2012) modified the item to make it shorter and to remove the use of the word “defective.” Ziegler administered both versions of the item to students and the percentage of students who answered correctly was higher for the original version in CAOS; however, the difference was not significant. As a result, the CAOS version of the assessment was included in BLIS-1. During the interviews, one student in particular

seemed to get lost in all of the words, so the decision was made to switch to the shorter version of the item for BLIS-2.

There was one particular item, Item 33, which threw off one of the more capable students who was able to correctly answer most of the other items. See Table 14 for a copy of the item that was in the BLIS-1 assessment. The item described a context where a dog was trained to smell stool samples to detect if a patient had bowel cancer. A p -value of less than .001 was reported and the student was expected to use the significance level of .05 to reject the null hypothesis. The student strongly believed that it was impossible for a dog to detect bowel cancer by smelling stool samples. As a result, he decided that there was a flaw in the design of the study and said “this is just a bogus experiment,” and did not use the p -value to make a decision. The student’s response was discussed with Dr. Joan Garfield, Dr. Michelle Everson, and one of the expert reviewers at the University of Minnesota. It was decided that because this response was from only one student, the item context, the dog attempting to detect bowel cancer, should be kept for the pilot administration of the assessment and student responses evaluated afterward. However, the question statement was changed to “Assuming the design of the experiment is good, use a significance level of .05 to make a decision” to try to prevent other students from making this mistake.

Table 14

Item 33 on the BLIS-1 Assessment that One Student Struggled with in the Cognitive Interview

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

New item based on a real-world context from Lock et al. (2013)

Item 12: 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 research question is “Does the dog correctly identify cancer more than half of the time?” The p-value is less than .001. Using a significance level of .05, what conclusion should be made? Explain why you chose to make your conclusion.

Part of Student 1’s response:

“The answer to the research question should be no because this is just a bogus experiment and there is nothing to say that the dog is detecting cancer. It could just be because the dog is detecting that particular individual.”

After all changes were made to the assessment, additional feedback was provided by one of the reviewers at the University of Minnesota, which led to minor changes in the wording of a couple of items. Appendix J1 includes all items in the BLIS-2 assessment and how they were changed from the BLIS-1 assessment.

Results from Pilot Test

There were two main purposes for conducting the pilot test: to see if the selected-response options were all viable options and to develop selected-response options for the constructed-response items. As mentioned previously, a total of 76 students enrolled in three introductory statistics courses completed the assessment. There were 12 students who were enrolled in a graduate-level course at the University of Minnesota, 31 students

enrolled in an undergraduate-level course at the University of Minnesota, and 33 students enrolled in an undergraduate-level course at Winona State University.

First, the 16 selected-response items were examined to see if the distractors were chosen by students. The percentage of students who chose each selected-response option in the pilot test was recorded (see Table 15). All distractors were chosen by at least one student for each item except for Item 4. In Item 4, the question described a study and the student was asked to select what type of study was conducted: observational, experimental, or survey. No students chose the survey option. It was decided that the option should still be kept because one of the reviewers emphasized that observational studies and experiments were not the only study designs, and that surveys should be included.

Table 15

Percentage of Students (N=76) Who Chose Each Selected-Response Option for the 16 Selected-Response Items Administered in the Pilot Test

Item	a	b	c	d	no response
3	59.2*	15.8	7.9	17.1	0.0
4	7.9	92.1*	0.0		0.0
5	82.9*	14.5	2.6		0.0
6	1.3	19.7	67.1*	11.8	0.0
10	6.6	7.9	11.8	72.3*	1.3
14	9.2	51.3*	35.5	1.3	2.6
15	7.9	10.5	18.4	60.5*	2.6
16	21.1*	22.4	27.6	26.3	2.6
21	7.9*	6.6	52.6	31.6	1.3
22	28.9	51.3*	18.4		1.3
27	43.4	42.1*	3.9	9.2	1.3
31	13.2	81.6*	3.9		1.3
34	34.2	61.8*	2.6		1.3
35	17.1*	11.8	7.9	61.8	1.3
36	90.8*	5.3	2.6		1.3
37	25.0	11.8	59.2*	2.6	1.3

Note. Items with no results presented for selected-response Option D represent an item that did not have an Option D. * indicates correct answer.

There were four items that had a low percentage of students who chose the correct answer. The percentage was determined to be low if the percentage of students who chose the correct option was lower than the percentage of students who chose a particular distractor. Considering the pilot was administered half-way through the semester, instructors were questioned to see if students had learned about the statistical content included in the four items. Item 16 measured the following learning outcome: understanding of the properties of standard deviation. The instructors for the two introductory statistics courses for undergraduate students had not taught students about standard deviation by the time students took the assessment, so the results were compared for the different courses (see Table 16). Students in the graduate-level course had learned

about standard deviation prior to taking the assessment and did better than students in the other courses. As a result, the item was not changed. For Item 21, which was about confidence intervals, only 8% of students answered correctly. According to instructors in all three courses, students had not learned about confidence intervals, so the item was not changed. Item 27 was about the logic of hypothesis testing and errors. A higher percentage of students chose the first distractor than the percentage of students who chose the correct response (see Table 17). Looking at the individual courses, two courses had a higher percentage of students who chose the correct response, so the item was not changed.

Table 16

Percentage of Students Who Chose Each Selected-Response Option for Items 16 and 27 Conditioned by Course

Course	a	b	c	d	no response	n
Item 16						
Undergraduate-level course ^a	18.2*	27.3	30.3	21.2	3.0	33
Undergraduate-level course ^b	16.1*	12.9	35.5	32.3	3.2	31
Graduate-level course ^b	41.7*	33.3	0.0	25.0	0.0	12
Total	21.1*	22.4	27.6	26.3	2.6	76
Item 27						
Undergraduate-level course ^a	48.5	36.4*	6.1	9.1	0.0	33
Undergraduate-level course ^b	41.9	38.7*	3.2	12.9	3.2	31
Graduate-level course ^b	43.3	40.0*	3.3	13.3	0.0	12
Total	43.4	42.1*	3.9	9.2	1.3	76
Item 35						
Undergraduate-level course ^a	12.9*	6.5	3.2	74.2	3.2	33
Undergraduate-level course ^b	21.1*	18.2	12.1	57.6	0.0	31
Graduate-level course ^b	41.7*	8.3	8.3	41.7	0.0	12
Total	17.1*	11.8	7.9	61.8	1.3	76

Note. * indicates correct answer.

^aCourse taught at Winona State University, ^bCourses taught at the University of Minnesota

Students performed poorly on Item 35. See Table 16 for the students' results and Table 17 for a copy of the item. Students did better in the graduate-level course compared with students in the undergraduate-level courses. The item was taken from the CAOS assessment, and in an analysis conducted by delMas et al. (2007), only 37.9% (N=715) of students answered the item correctly on a posttest, which is comparable to students in the graduate-level course in this study. Results from the CAOS assessment also indicated a fairly high item-total correlation (Robert delMas, personal communication, April 5, 2014). An updated version of the item that mentioned that the students were selected

randomly was included in an early version of the GOALS assessment. It was decided to include this updated version of the item in the BLIS-3 assessment.

Table 17

Item 35 on the BLIS-2 Assessment

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

Item was taken from the CAOS assessment, Item 38

Item 35: A college official conducted a survey 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. Which of the following does **NOT** affect the college official's ability to generalize the survey results to all dormitory students?

- 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.
-

For the 21 constructed-response items, minor changes were made to the item stems, and selected-response options were written. Item stems that asked students to explain their answers were changed so the items no longer requested explanations. Students' responses were put into groups of similar responses. To write the correct selected-response options, students' correct answers were examined and responses that were most complete and were related to the specified learning outcome were chosen. For most items, multiple correct answers were chosen and merged to create a well-written selected-response option. Most selected-response options were also reworded to be similar in language, length, and style as recommended by Haladyna and Rodriguez (2013).

For many items, the most common student responses were used to create the distractors. A description of the distractors that were created, based on considerations

other than using the most common student responses, are described here. For the first item, students were asked to define the sample and the population. There were no students who flipped the sample and population, but it was used as a distractor because there were not many other student responses that would make good distractors. Many students only defined the sample and others defined only the population, which limited the student responses that could be used to create distractors. Item 7 was very similar. Students were asked to define the statistic and parameter of interest for a particular situation. Again, no students flipped the statistic and parameter, but it was added as a distractor.

There were multiple items that were similar or seemed to naturally fall into groups, but these items did not have many incorrect student responses. Item 8 asked students to describe why two samples did not produce the same mean. There were 11 students who gave an incorrect answer related to sample size, and this was used as one of the distractors. The other incorrect responses did not appear to share anything in common. Therefore, one student's response that mentioned representativeness was used because the other responses seemed nonsensical. For Item 9, students had to interpret a dotplot. A handful of students used anecdotal information to answer the question, and only one student provided an answer that was statistically incorrect; the student referred to the mode as the majority. Considering there were no other statistically incorrect responses, that student's response was used to create one of the distractors. Item 13 asked students to provide an interpretation of a sample mean. The most common incorrect answer was that students referred to the majority in their interpretation. There was only

one student who interpreted it as the median, but, considering there were no other common responses, the interpretation referring to the median was used as a distractor.

One item writing recommendation is to avoid giving clues to the right answer (Haladyna & Rodriguez, 2013). In Item 23, students were asked a yes or no type of question. Students were presented with a randomization distribution and observed result, and were asked if there was evidence against the null hypothesis. The correct answer is *yes*, because the proportion of re-randomized sample mean differences equal to or above the sample statistic is very small. There were two groups of common incorrect responses which were used to make distractors. One of them included an incorrect answer of *no* and the other included an incorrect explanation for choosing *yes*. If those were the only distractors included, students would see one selected-response option for the choice of *no* and two selected-response options for the choice of *yes*. Having two selected-choice options for choosing *yes* could clue students into thinking the correct answer is *yes*. Therefore, other responses that included *no* as the answer were examined. Two students justified a response of *no* by stating that the proportion of re-randomized sample mean differences equal to or above the sample statistic is very small. Therefore, that response was considered to be a reasonable misconception and was added as a distractor.

In Item 26, students were asked to compute a p -value for a one-tailed test (see Table 18). In the pilot test, students were asked to explain how they found their p -value. Multiple students reported a p -value of .04; however, two different explanations were reported. Some students arrived at the answer of .04 by taking the number of re-randomized sample mean differences at least as large as the observed statistic and incorrectly dividing by the sample size instead of dividing by the number of trials. Other

students arrived at the answer of .04 by computing a two-tailed p -value instead of a one-tailed p -value. In order to create distractors that would separate the two incorrect responses, the randomization distribution plot was changed to include three re-randomized sample mean differences larger than the observed statistic. Therefore, if a student made the first incorrect mistake, their p -value would be .06 and if a student made the second mistake, their p -value would be .05.

Table 18

Item 26 on the BLIS-2 Assessment and Student Explanations for Incorrectly Reporting a P-value of .04

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

New item based on a real-world context from Steinberg, Levine, Askew, Foley, and Bennett (2013)

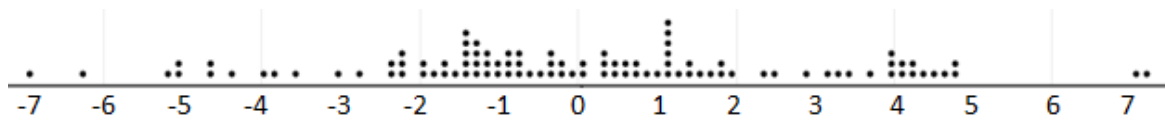
Items 25 and 26 refer to the following situation:

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. (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)

Item 26: Researchers are interested in whether 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' interest. Explain how you found his *p*-value so someone else can replicate your work.

Student responses:

Divided by sample size instead of number of trials: “p-value=0.04. I found this p-value by dividing 2 by 50 because there were two dots located above 5.4 and there were 50 women in the study.”

Computed a two-tailed *p*-value instead of a one-tailed *p*-value: “the p value is about .05. I only found 4 values out of the 100 to be more extreme”

One item, Item 29, was deleted from the assessment (see Table 19). Looking at the student responses, it was clear that the item was not measuring the intended learning outcome: understanding that every model is based on assumptions that limit our scope of inferences. Students' responses indicated that Item 29 came closer to measuring the following learning outcome: understanding of the factors that allow a sample of data to be generalized to the population. Considering that an item was already included in the assessment related to the ability to generalize, the item was deleted from the assessment.

Table 19

Item 29 on the BLIS-2 Assessment Which Did Not Appear to Measure the Intended Learning Outcome

Learning Outcome: Understanding that every model is based on assumptions which limit our scope of inferences

New item based on a real-world context from delMas (2013)

Item 29: Researchers are interested in whether 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' interest. Explain how you found his p -value so someone else can replicate your work.

Student responses:

"No because you are only using voters who have telephones. You can not say the sample represents the U.S. population."

"No. They did not take into account the number of people who do not vote at all."

"No, they are only surveying people from 100 residential phone numbers in each state. To get more accurate results, they would need a lot larger sample size."

"Yes because it is a random sample."

Item 30 provided students with a research question and asked them what the null hypothesis and alternative hypothesis were. A majority of students wrote the hypothesis statements in the form of a sentence and not using symbols. Common mistakes included switching the null hypothesis and alternative hypothesis, and not referring to the parameter. Including both the null hypothesis and the alternative hypothesis in each

selected-response option led to an item that became quite lengthy, so the original question was split into two questions. The first question asked students to select the correct null hypothesis statement and the second question asked students to select the correct alternative hypothesis statement.

After the selected-response options were written, the items were then examined by one of the reviewers at the University of Minnesota and Dr. Michelle Everson. Changes in wording were made and a few distractors were added. For example, it was suggested that for Item 13, a distractor should be included with an incorrect interpretation of a sample mean that referred to the population. In another item, students were asked to interpret a percentage and it was noted that a common misconception is that students tend to think about individuals rather than think about the sample as a whole. There was one student who wrote about individuals, so that response was added as a distractor.

After all changes were made, the BLIS-3 assessment was complete. See Appendix K for a copy of the changes made to the BLIS-2 assessment. Also, considering Item 29 was deleted and Item 30 was split into two items, the test blueprint needed to be slightly modified. The learning outcome for Item 29 was deleted and the additional item number was added to the learning outcome for the original Item 30. See Table A4 in Appendix A for a copy of the BLIS Test Blueprint-3.

The instructor survey was also piloted with the instructors who administered the assessment to their students. Although no problems were found with the survey, changes were made to the survey before administering it in the field test. First, one question asked instructors to report the state that their institution was in, and the question was changed to ask what country their institution was in. Questions were also added to determine which

simulation-based methods instructors included in their courses and to learn about what value instructors perceived the BLIS assessment to have for statistics educators.

Results from the Field Test

This sub-section reports the results from the field test including descriptive statistics, as well as evidence of reliability, validity, and value.

Descriptive statistics. First, the percentage of students, out of 940, who chose each selected-response option for each item, was computed (see Table 20). All items, except for Items 21 and 35, had the highest percentage of students chose the correct option. Both Items 21 and 35 had a low percentage correct on the pilot test as well as the field test.

Table 20

Percentage of Students (N=940) Who Chose Each Selected-Response Option for All 37 Items Administered in the Field Test

Item	a	b	c	d
1	11.7	6.9	81.4*	
2	6.2	7.9	49.9*	36.1
3	56.9*	16.9	3.6	22.6
4	9.9	89.6*	0.5	
5	90.0*	9.1	0.9	
6	1.1	12.4	76.8*	9.7
7	32.2	14.7	11.1	42.0*
8	40.0	18.7	41.3*	
9	70.3*	26.4	3.3	
10	9.3	7.3	8.3	75.1*
11	23.9	36.9	39.1*	
12	13.9	21.1	3.0	62.0*
13	44.7*	31.4	6.2	17.8
14	4.0	48.0*	47.0	1.0
15	5.1	4.4	14.0	76.5*
16	37.4*	15.3	27.0	20.2
17	17.7	12.3	70.0*	
18	27.4	53.0*	19.6	
19	4.7	18.2	16.8	60.3*
20	16.2	27.4	48.7*	7.7
21	19.9*	4.3	39.0	36.8
22	27.3	64.7*	8.0	
23	10.4	23.2	49.3*	17.1
24	61.3*	14.3	24.5	
25	53.4*	21.6	25.0	
26	53.3*	35.0	11.7	
27	41.3	46.6*	5.9	6.3
28	9.7	17.4	72.9*	
29	12.7	10.3	64.3*	12.8
30	5.9	17.4	15.4	61.3*
31	20.6	71.8*	7.6	
32	12.7	12.6	63.5*	11.3
33	63.4*	27.3	9.3	
34	25.9	70.9*	3.3	
35	27.1*	11.4	11.5	50.0
36	89.1*	8.5	2.3	
37	19.7	11.1	64.0*	5.2

Note. Only students who completed the entire assessment are included in this table. Items with no results presented for selected-response Option D represent an item that did not have an Option D. * indicates correct answer.

The BLIS-3 assessment contained five testlets, which were pairs of items that included a common stem (Downing, 2006a). One testlet included two items, Items 29 and 30, that shared the same learning outcome: ability to determine a null and alternative hypothesis statement based on a research question. Student scores were examined for Item 29 and Item 30. Recall that Item 29 asked students to provide the null hypothesis for a particular research question and Item 30 asked students to provide the alternative hypothesis for the same research question. Table 21 includes the results for these two items. A majority of students either answered both items correctly or both items incorrectly. Therefore, it was decided that these items would be scored together as one item (Item 29/30) for the remainder of the analyses presented in this chapter. So, instead of examining 37 item scores, 36 item scores were examined. Students who answered one or both Items 29 and 30 incorrectly received a score of 0 and students who answered both items correctly received a score of 1. Partial credit was not given if students answered only one item correctly.

Table 21

Total Percentage of Students Who Answered Correctly or Incorrectly for Items 29 and 30 on the BLIS-3 Assessment

		Item 29	
		Incorrect	Correct
Item 30	Incorrect	31.1	4.7
	Correct	7.7	56.6

The other four testlets had item pairs with different learning outcomes. Therefore, each item was scored separately. Testlets have been claimed to create local dependence for items within a testlet (Downing, 2006a), so testlet scores were also created. Each testlet was scored as a 0, 1, or 2, where a score of 0 indicated that both items in the testlet

were incorrect and a score of 2 indicated that both items in the testlet were correct. Statistical analyses presented in the remainder of this chapter were conducted without testlet scores and compared with analyses conducted with testlet scores.

Students' total scores were examined for the 36 items. The average total score was 21.41 out of 36 ($s=6.25$, $N=940$), or 59.5%. See Figure 6 for a visual representation of students' total scores for the 36 items.

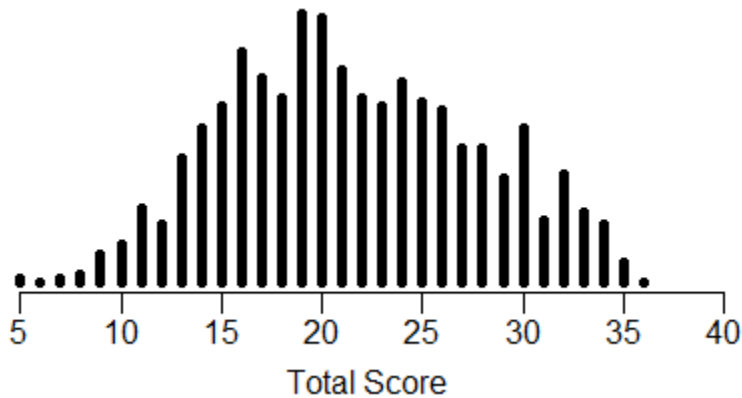


Figure 6. Dotplot of students' total scores for the 36 items on the BLIS-3 assessment.

Reliability. Coefficient alpha was computed to check for the overall reliability of the BLIS-3 assessment at the test level. When examining the 36 items scores individually for students who completed the entire assessment, coefficient alpha was .83. The Spearman-Brown formula was used to predict what coefficient alpha would be for an assessment with 32 items rather than an assessment with 36 items. The predicted coefficient alpha was .81. When testlet scores were incorporated, coefficient alpha was actually .82. The small differences suggest that there is not local dependence for items in testlets.

Assumptions for analysis based on IRT. In order to conduct analysis based on IRT, the assumptions of unidimensionality and local independence were examined using CFA. Two scree plots of eigenvalues were examined: one for the BLIS-3 assessment that

consisted of the 36 individual item scores, and the second for the BLIS-3 assessment that consisted of 32 item and testlet scores. Both scree plots of eigenvalues showed evidence that the BLIS-3 assessment consisted of one factor, because the eigenvalues of the factors leveled off after the first factor (see Figure 7). The scree plot that included testlet scores, however, looked slightly better because there was slightly less of a jump between the third and fourth factors compared with the scree plot that did not include testlet scores. Therefore, the scree plots show evidence of unidimensionality and possibly local dependence for items in testlets.

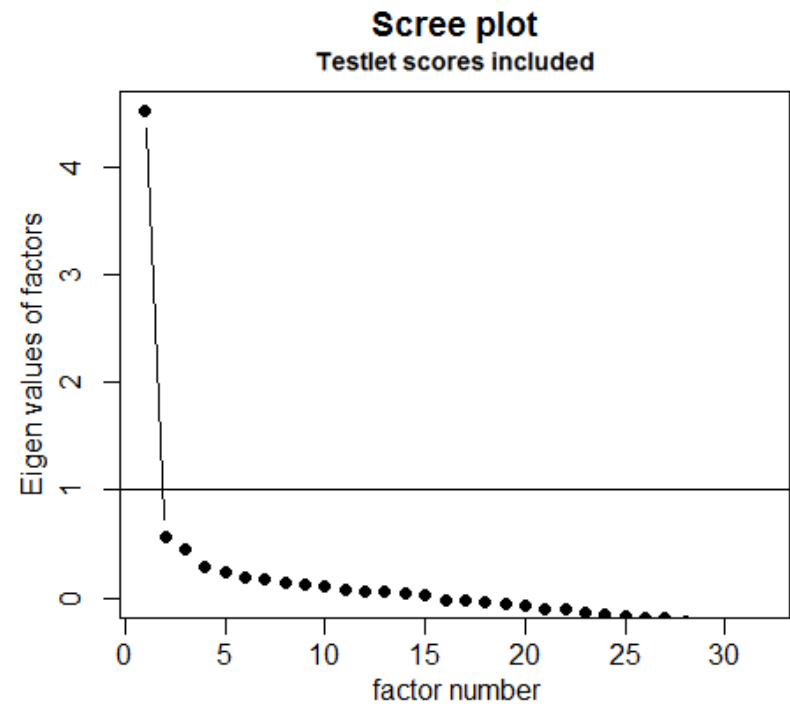
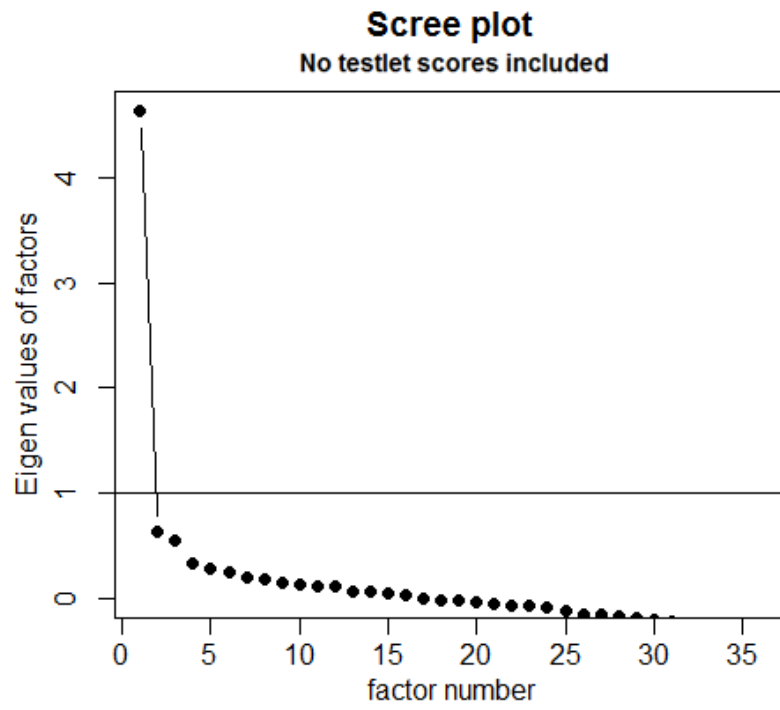


Figure 7. Scree plots of eigenvalues for the BLIS-3 assessment.

Further evidence of unidimensionality was collected by examining the factor loadings and model fit indices. Each of the factor loadings were positive, indicating good fit for a single-factor model (see Table 22). In addition, all of the model fit indices examined, except for the chi-squared goodness-of-fit statistic, showed good fit (see Table 23). According to Bentler and Bonett (1980), when sample sizes are too large, the chi-square goodness-of-fit test will reject almost any model as inadequate. Therefore, other fit indices were examined: the Tucker-Lewis Index (TLI), Bentler's Comparative Fit Index (CFI), and the Root Mean Square Error of Approximation (RMSEA). Hu and Bentler (1999) recommended a cutoff value close to .95 or higher for the TLI and CFI, and a cutoff value close to .06 or lower for the RMSEA.

Table 22

Factor Loadings for One-Factor CFA Model with 36 Individual Items

Item	Loading	SE	<i>p</i> -value	Item	Loading	SE	<i>p</i> -value
1	0.477	0.037	0.000	19	0.518	0.037	0.000
2	0.321	0.041	0.000	20	0.607	0.034	0.000
3	0.441	0.038	0.000	21	0.376	0.058	0.000
4	0.339	0.049	0.000	22	0.296	0.041	0.000
5	0.421	0.047	0.000	23	0.410	0.039	0.000
6	0.329	0.043	0.000	24	0.552	0.033	0.000
7	0.571	0.037	0.000	25	0.427	0.039	0.000
8	0.084	0.045	0.061	26	0.524	0.035	0.000
9	0.350	0.041	0.000	27	0.492	0.039	0.000
10	0.389	0.041	0.000	28	0.549	0.036	0.000
11	0.206	0.044	0.000	29/30 ^a	0.779	0.025	0.000
12	0.186	0.043	0.000	31	0.499	0.035	0.000
13	0.354	0.042	0.000	32	0.475	0.036	0.000
14	0.285	0.042	0.000	33	0.551	0.035	0.000
15	0.509	0.037	0.000	34	0.494	0.035	0.000
16	0.603	0.039	0.000	35	0.559	0.044	0.000
17	0.583	0.033	0.000	36	0.596	0.036	0.000
18	0.493	0.037	0.000	37	0.556	0.034	0.000

^aItems 29 and 30 were combined to make one item score

Table 23

Fit Indices for One-Factor CFA Models

Fit indices	No testlet scores	Testlet scores
χ^2	535.936	462.747
CFI	0.944	0.952
TLI	0.961	0.968
RMSEA	0.027	0.027

In order to check if the remaining four testlets resulted in a violation of the local independence assumption, the tetrachoric correlation residuals were examined (see Appendix L). The correlation residuals for the item pairs in testlets were similar in magnitude compared to other item pairs. Another single-factor CFA was run taking into account the items that were in testlets. The new factor loadings were also positive indicating good fit for a single-factor model (see Table 24). The model fit indices indicated that the single-factor CFA model fit slightly better when incorporating testlet scores (see Table 23). Therefore, it was decided that including testlet scores was acceptable to meet the local independence assumption.

Table 24

Factor Loadings for One-Factor CFA Model with 32 Individual Items and Four Testlets

Item/ testlet	Loading	SE	<i>p</i> -value	Item/ testlet	Loading	SE	<i>p</i> -value
1	0.476	0.037	0.000	18/19	0.603	0.028	0.000
2	0.316	0.041	0.000	20	0.605	0.034	0.000
3	0.441	0.038	0.000	21	0.378	0.057	0.000
4	0.341	0.050	0.000	22	0.296	0.041	0.000
5/6	0.425	0.037	0.000	23/24	0.582	0.030	0.000
7	0.570	0.036	0.000	25/26	0.559	0.029	0.000
8	0.086	0.045	0.054	27	0.492	0.039	0.000
9	0.350	0.041	0.000	28	0.549	0.037	0.000
10	0.387	0.041	0.000	29/30 ^a	0.780	0.025	0.000
11	0.205	0.044	0.000	31	0.498	0.036	0.000
12	0.187	0.043	0.000	32	0.476	0.036	0.000
13	0.355	0.041	0.000	33	0.551	0.035	0.000
14	0.283	0.042	0.000	34	0.493	0.035	0.000
15	0.510	0.037	0.000	35	0.556	0.043	0.000
16	0.602	0.039	0.000	36	0.593	0.037	0.000
17	0.583	0.033	0.000	37	0.554	0.034	0.000

^aItems 29 and 30 were combined to make one item score

IRT models. Three IRT models were fit to the data: the Rasch model (Hambleton et al., 1991), 2 parameter logistic (2PL) model (Hambleton et al., 1991), and partial credit (PC) model (Masters, 1982). Fit indices indicated that the PC model fit the data best (see Table 25). To examine the fit of the PC model in more detail, a parametric bootstrap approximation to the Pearson chi-squared goodness-of-fit measure provided evidence that the PC model had good fit ($p = .27$). Therefore, the PC model was used in the remaining analysis.

Table 25

Fit Indices for Rasch, 2PL, and PC Models

Fit indices	Rasch	2PL	PC
AIC	39075.86	38661.80	37121.64
BIC	39255.15	39010.70	37296.09
Log Likelihood	-19500.93	-19258.90	-18524.82

Precision. Using the PC model, item information and test information were examined. Information is used to understand how well an assessment discriminates between students of different ability levels. The item information curves provide reliability evidence for individual items and display when particular items are useful in differentiating among individuals. Examining the item information curves in Figure 8, it can be seen that there are items that differentiate individuals across abilities levels indicating that the assessment should perform well in estimating ability levels ranging from approximately -2.5 to 2.5. There are slightly more items measuring students with lower ability levels than items for higher ability levels.

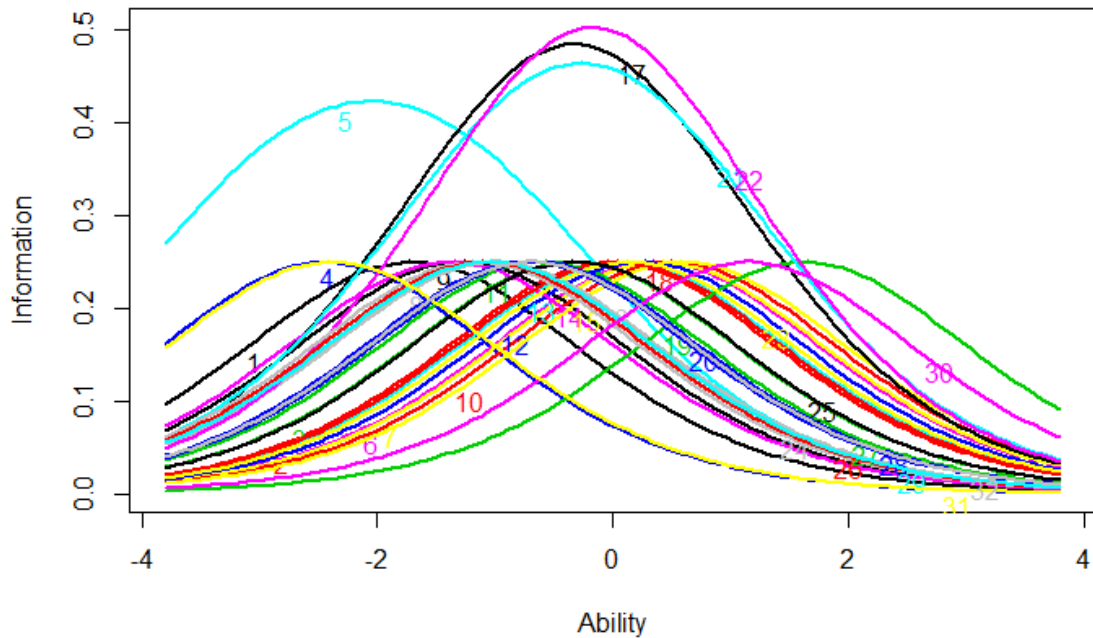


Figure 8. Item information curves for the 32 items and 4 testlets.

The test information function is presented in Figure 9. Similar to what was seen in the item information curves, it appears that the BLIS-3 assessment provides information for students at lower and higher abilities; however, the assessment is slightly better at estimating students' abilities at the lower end. From a different perspective, the standard error of measurement curve displays that there is less error in estimating students' ability levels that are closest to 0, where the test information is highest. The measurement error is highest for students at higher ability levels, indicating the lowest amount of information.

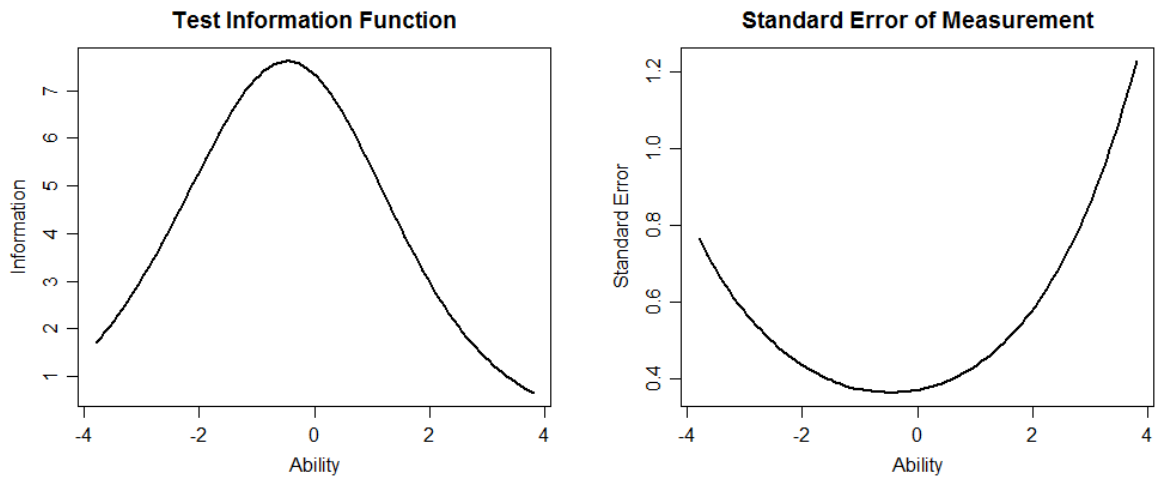


Figure 9. Test information function and standard error of measurement for the BLIS-3 assessment.

Validity. The difficulty estimates and standard errors were examined to collect further evidence of validity (see Table 26). Item difficulties ranged from the least difficult at -3.058 to the most difficult at 1.619. There are more items with lower difficulties than items with higher difficulties indicating that new items are needed to measure students at higher ability levels.

Table 26

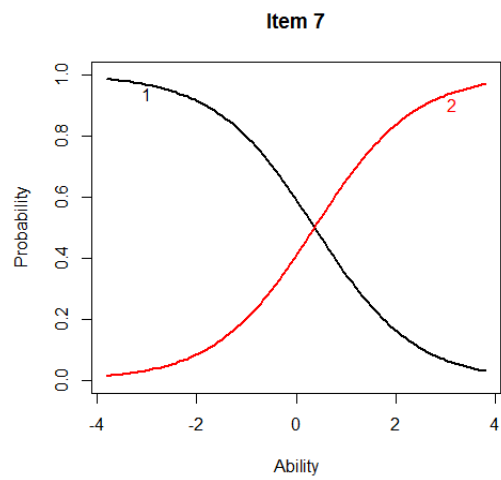
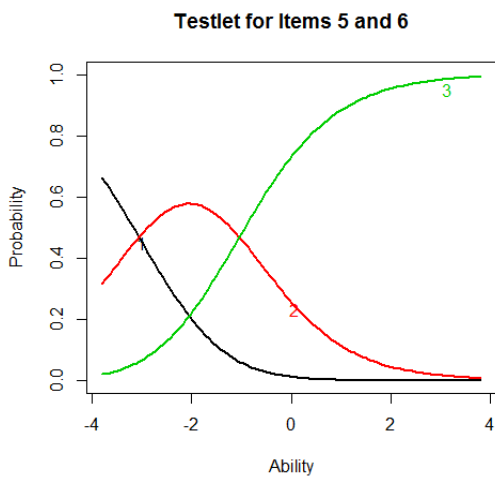
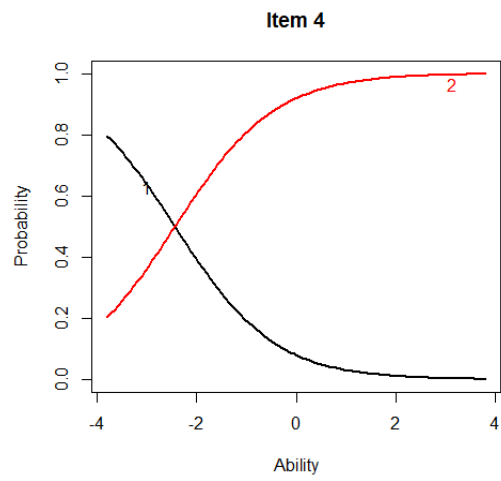
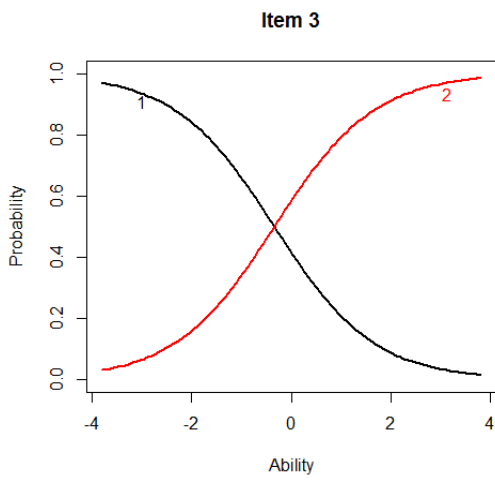
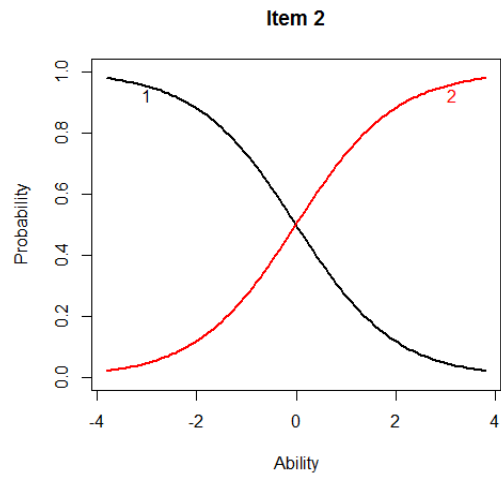
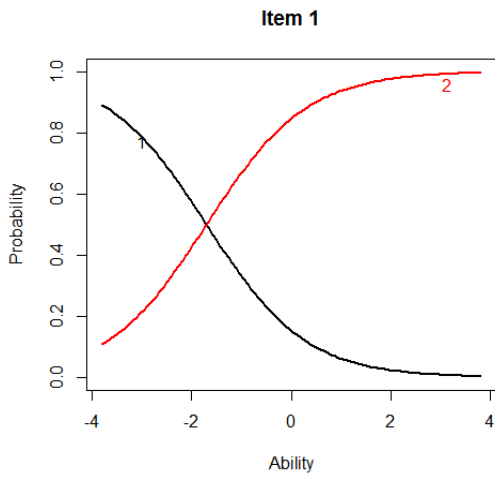
Item Parameters for the PC Model with 32 Individual Items and four Testlets

Item/ testlet	Difficulty 1 (SE)	Difficulty 2 (SE)	Item/ testlet	Difficulty 1 (SE)	Difficulty 2 (SE)
1	-1.704 (0.094)		18, 19	-1.093 (0.096)	0.424 (0.086)
2	-0.009 (0.078)		20	0.046 (0.078)	
3	-0.341 (0.079)		21	1.619 (0.094)	
4	-2.438 (0.115)		22	-0.721 (0.080)	
5, 6	-3.058 (0.216)	-1.045 (0.085)	23, 24	-1.112 (0.095)	0.566 (0.087)
7	0.366 (0.079)		25, 26	-0.858 (0.092)	0.514 (0.088)
8	0.402 (0.079)		27	0.147 (0.078)	
9	-1.016 (0.083)		28	-1.159 (0.085)	
10	-1.291 (0.087)		29/30 ^a	-0.326 (0.078)	
11	0.506 (0.080)		31	-1.099 (0.084)	
12	-0.588 (0.080)		32	-0.662 (0.080)	
13	0.238 (0.079)		33	-0.657 (0.080)	
14	0.081 (0.078)		34	-1.045 (0.084)	
15	-1.375 (0.088)		35	1.150 (0.086)	
16	0.591 (0.080)		36	-2.390 (0.114)	
17	-0.999 (0.083)		37	-0.688 (0.080)	

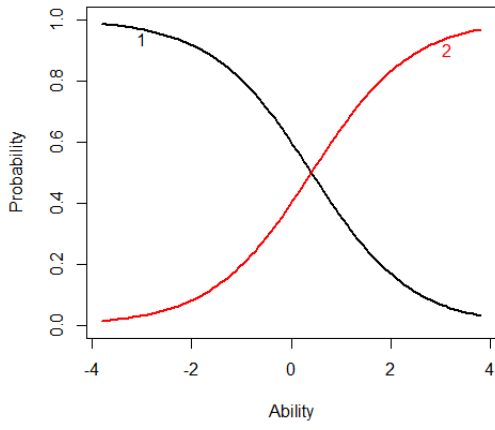
^aItems 29 and 30 were combined to make one item score

Item Characteristic Curves were created for each item or testlet (see Figure 10).

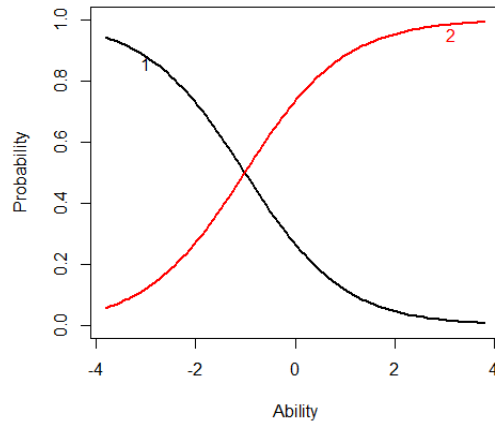
Multiple items discriminate between students at lower ability levels (e.g., Item 1, testlet for Items 5 and 6) and multiple items discriminate between students with ability levels close to 0 (e.g., Item 2, testlet for Items 18 and 19). Items 21 and 35 appear to discriminate between students at higher ability levels.



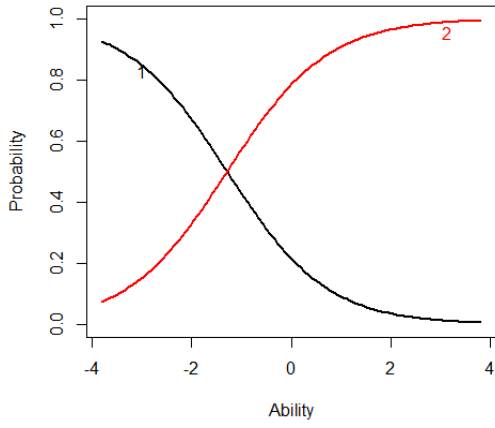
Item 8



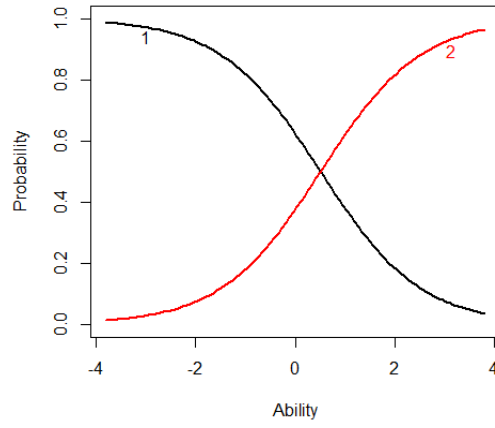
Item 9



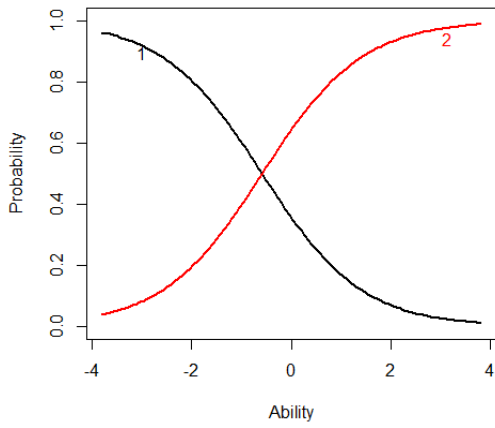
Item 10



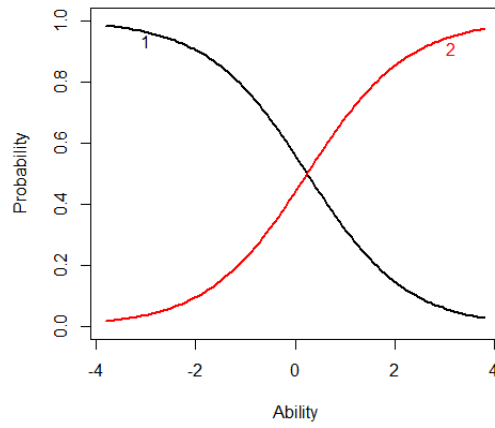
Item 11



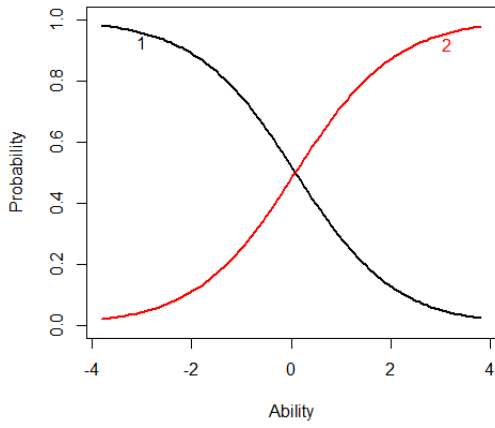
Item 12



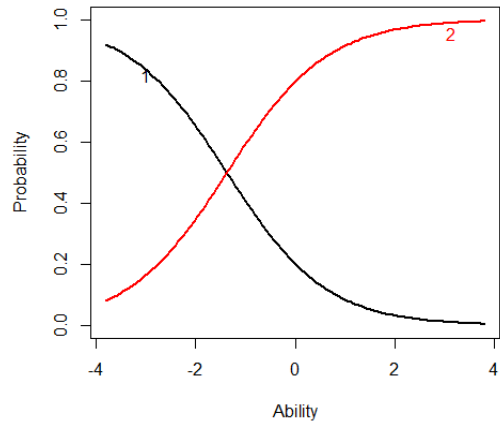
Item 13



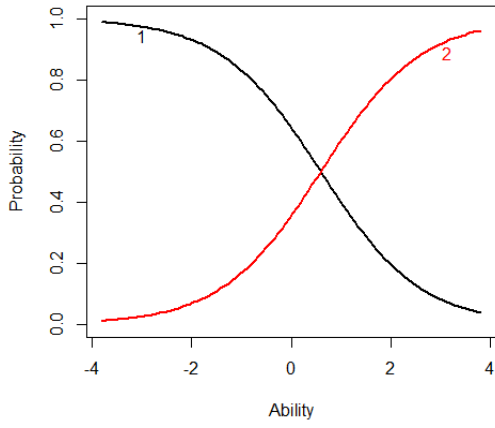
Item 14



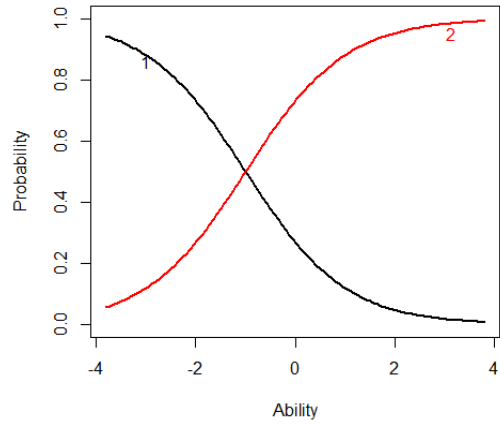
Item 15



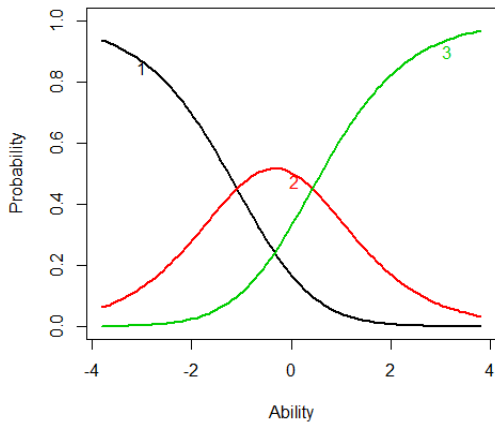
Item 16



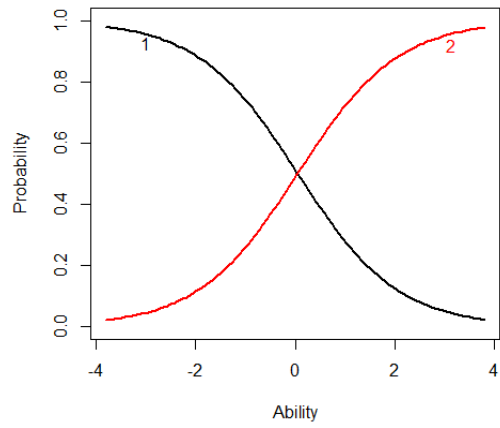
Item 17



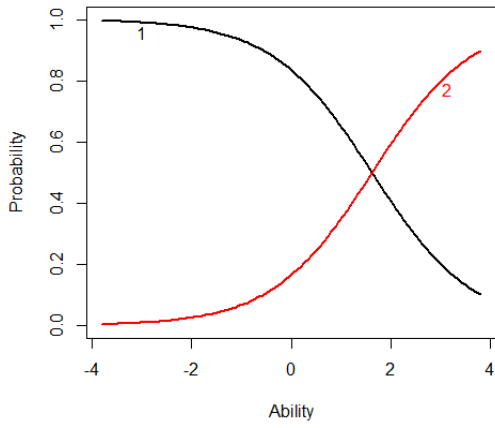
Testlet for Items 18 and 19



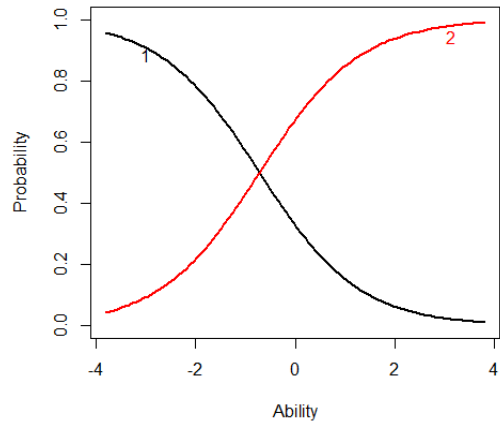
Item 20



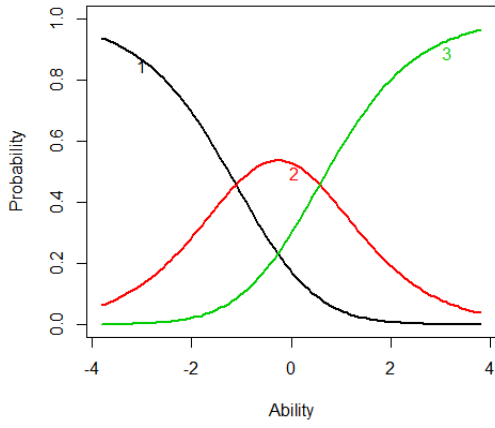
Item 21



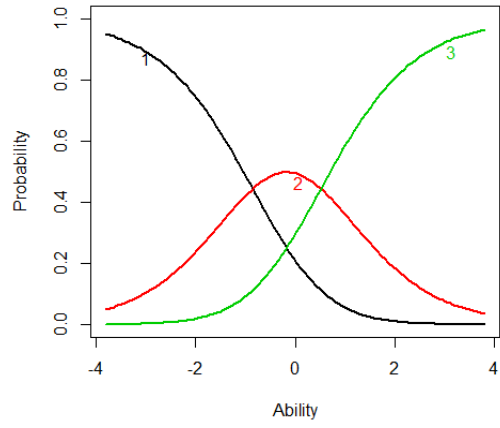
Item 22



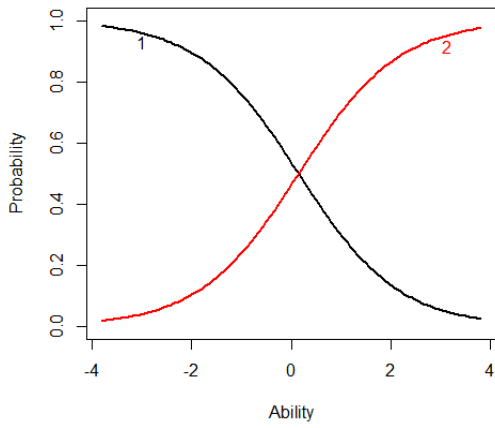
Testlet for Items 23 and 24



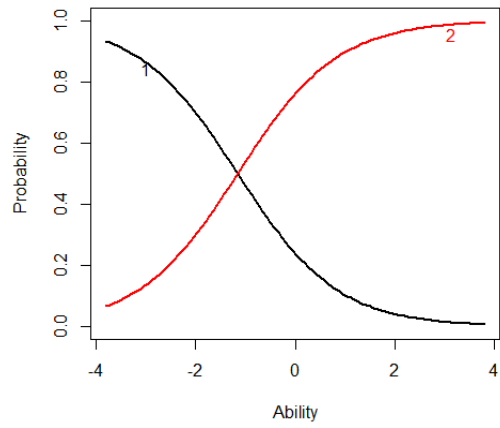
Testlet for Items 25 and 26



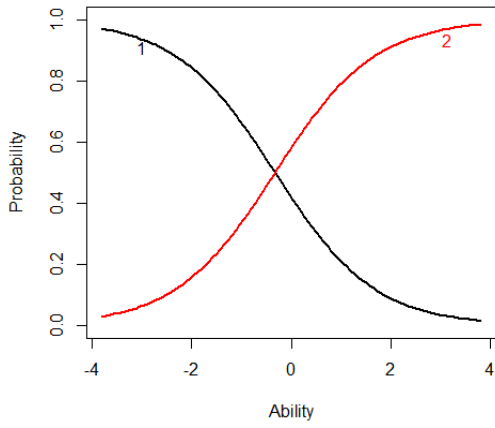
Item 27



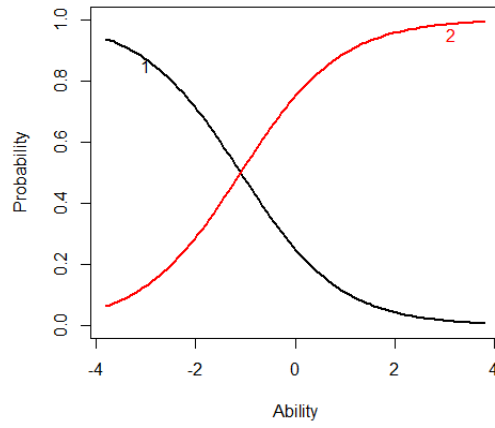
Item 28



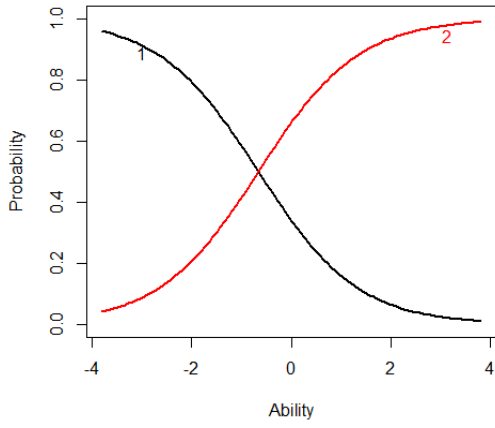
Item 29/30



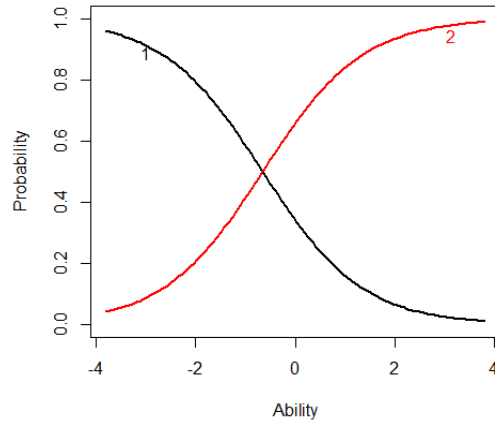
Item 31



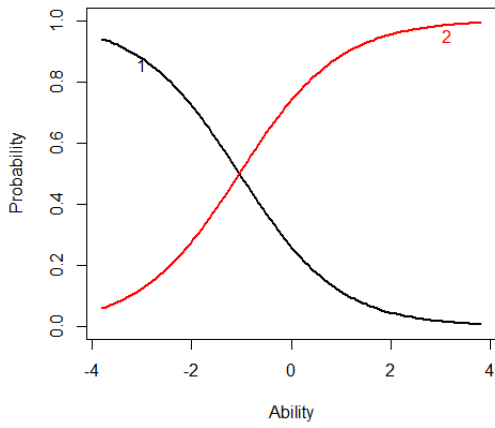
Item 32



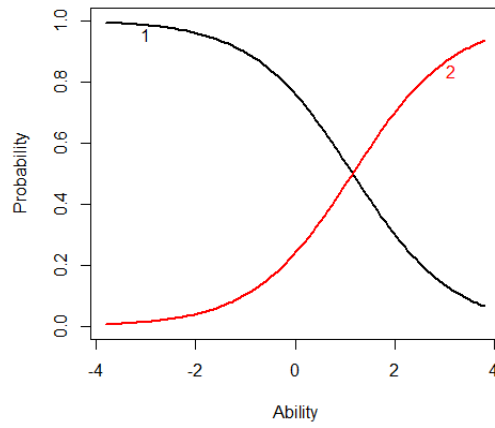
Item 33



Item 34



Item 35



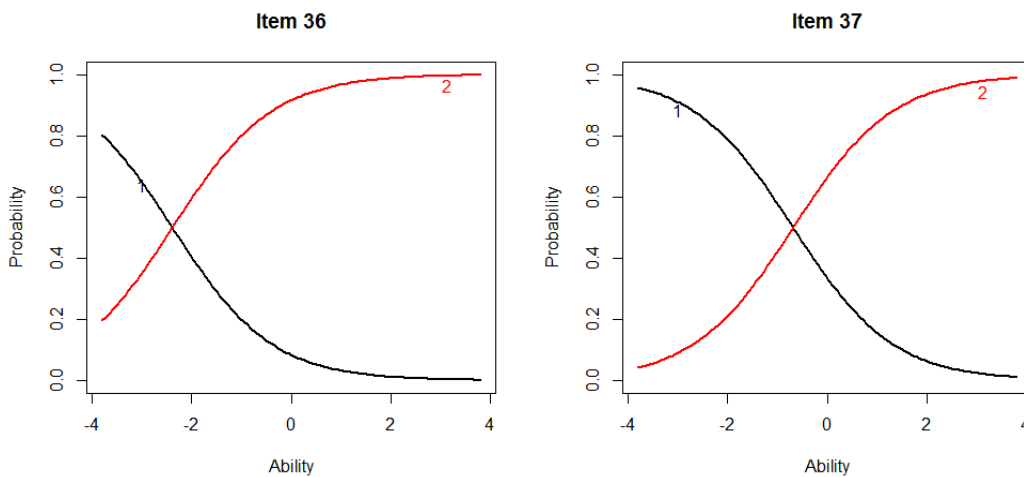


Figure 10. Item characteristic curves of 28 items and 4 testlet-based items. For the 28 items, a value of 1 represents an incorrect response and a value of 2 represents a correct response. For the 4 testlet-based items, a value of 1 represents an incorrect response, a value of 2 represents a partially correct response, and a value of 3 represents a correct response.

Value. Evidence of the value of the BLIS assessment to statistics educators was collected from instructors who administered the BLIS-3 assessment to their students. The value to instructors and value to educational researchers were rated separately on a 4-point scale from “Not at all valuable” to “Very valuable.” See Table 27 for the survey results. Two statistics instructors filled out the survey twice because they administered the assessment in two different courses. One of the two instructors gave the same value ratings, so his results were included in the analysis. The other instructor gave different ratings, so his results were not used. All statistics instructors who participated believed the BLIS assessment provides value to statistics education researchers and statistics instructors. It appears that statistics instructors believed the BLIS assessment is more valuable to statistics education researchers than to statistics instructors; however, both were rated high.

Table 27

Value Ratings Reported by 26 Introductory Statistics Instructors

	Counts (Percentages)			
	Not at all valuable	A little valuable	Valuable	Very valuable
Instructors	0 (0.0)	3 (11.5)	18 (69.2)	5 (19.2)
Educational researchers	0 (0.0)	0 (0.0)	14 (53.8)	12 (46.2)

Chapter 5

Discussion

This chapter includes a synthesis of the results from the study. Conclusions about the reliability, validity, and value of the BLIS assessment are provided. A reconceptualization of statistical literacy is offered in light of what was learned in conducting this study. The chapter concludes with limitations of the study and implications for teaching and for future research.

Summary of the Study

The purpose of this study was to develop a new assessment of statistical literacy (called the BLIS assessment) to be used in an introductory statistics course that incorporates, to some extent, simulation-based methods. Multiple steps were taken to collect evidence of reliability, validity, and value of the BLIS assessment. Data was collected from expert reviewers, students, and introductory statistics instructors.

The development of the BLIS assessment began with a test blueprint. The preliminary test blueprint was created based on a review of textbooks (Gould & Ryan, 2013; Catalysts for Change, 2013; Lock et al., 2013; Tintle et al., 2013). Expert reviews were conducted with six statistics education experts who rated the importance of each learning outcome to be included on the assessment. Based on the feedback of expert reviewers, changes were made to the preliminary test blueprint to create the BLIS Test Blueprint-1.

The preliminary BLIS assessment was created based on the test blueprint. Nineteen items were taken from existing instruments: the CAOS test (delMas et al.,

2007), ARTIST Topic Scale tests (Garfield et al., 2002), ARTIST item database (Garfield et al., 2002), and an early version of the GOALS assessment (Garfield et al., 2012). In addition, 18 new items were written to measure the learning outcomes.

The assessment was modified three times after the preliminary version was created. Expert reviewers rated how well each item on the preliminary version measured the specified learning outcome, and their feedback was used to modify items for the BLIS-1 assessment. Cognitive interviews were conducted with students to make the next round of changes for the BLIS-2 assessment. The last round of changes took place after the assessment was administered in a small-scale pilot test.

The BLIS-3 assessment was administered in a large-scale field test with students enrolled in an introductory statistics course. Analyses based on CTT and IRT were conducted to further examine reliability and validity. Instructors who administered the assessment also completed a survey that was designed not only to gather demographic information about the courses, but to gather evidence of the value of the BLIS assessment.

Synthesis of the Results

Reliability. Reliability evidence was collected through expert reviews, cognitive interviews with students, a small-scale pilot test, and a large-scale field test. As noted earlier, low reliability can be the result of measurement error, and all assessments have measurement error due to natural variability (Weathington et al., 2010). Two sources of measurement error that can be minimized are instrument error and participant variability.

Instrument error includes issues in wording and organization (Weathington et al., 2010). One source of participant variability that could be the result of instrument error is the misunderstanding of items. Instrument error and participant variability were examined through the expert reviews of the assessment, cognitive interviews with students, and the small-scale pilot test. In the expert reviews of the assessment, reviewers provided suggestions about how to word items differently in order to better measure the intended learning outcomes. During the interviews, students completed the assessment while reading the items aloud and verbalizing their thoughts. Students were able to point out small issues in the wording of the items. The interviews also provided insight about whether the language of the items were leading students down the wrong path, and this allowed additional changes to be made. The BLIS-2 assessment used in the pilot test included 37 items, 21 of which were constructed-response items. Students' written responses brought to light additional minor changes to be made in the wording of two items. In addition, one item was deleted because it was clear that it was not measuring the intended learning outcome.

Reliability evidence was also collected in the large-scale field test of the BLIS-3 assessment. Coefficient alpha was high when testlets were taken into account as well as when they were not; .83 and .82 respectively. It appears that there is a slight benefit to incorporating testlet scores in the analyses. While no evidence of item dependence appeared when examining the tetrachoric correlation residuals for the single-factor CFA without testlet scores, the single-factor CFA model fit slightly better when incorporating testlet scores based on the CFI and TLI model fit indices. Furthermore, when examining

IRT models, the PC model, which incorporated testlet scores, fit better than the Rasch model and 2PL model, which did not incorporate testlet scores.

Examining the item information functions, test information function, and standard error of measurement function from the PC model, it appears that the precision in estimating students' abilities is highest when their abilities are closest to 0. Also, precision is higher for students with lower ability levels than students with higher ability levels because there are more items with low difficulty levels than items with high difficulty levels.

Considering the precautions taken in the wording of items and the results from the statistical analysis, the BLIS-3 assessment appears to have high reliability. Many changes in the wording of items on the preliminary version of the assessment were made based on the expert reviews, and in each of the following versions of the assessment, fewer changes were needed based on student data. Only a handful of items in the BLIS-2 assessment required changes in wording. This suggests that the wording of the items in the BLIS-3 assessment is of high quality.

Validity. Multiple pieces of evidence were gathered to create a validity argument for the reasonableness of inferences. The BLIS assessment was developed to make inferences about students' statistical literacy in an introductory statistics course that incorporates, to some extent, simulation-based methods. Validity evidence was collected through expert reviews, cognitive interviews with students, a small-scale pilot test, and a large-scale field test.

The first consideration in collecting evidence of validity was to ensure the BLIS assessment was measuring important aspects of statistical literacy. First, in order to choose relevant topics and learning outcomes for statistical literacy, textbooks used in introductory statistics courses that include simulation-based methods were reviewed (Gould & Ryan, 2013; Catalysts for Change, 2013; Lock et al., 2013; Tintle et al., 2013). The preliminary test blueprint was based on the textbook review, which was then reviewed by six expert reviewers. The reviewers rated the importance of each learning outcome to be included on the assessment. The learning outcomes that were rated highest were included in the final version of the BLIS test blueprint. Furthermore, the reviewers provided feedback on what they felt was missing from the test blueprint, and this resulted in four new learning outcomes.

Multiple steps were taken to provide evidence that the BLIS assessment was measuring one construct, statistical literacy. The test blueprint was developed to include important learning outcomes of statistical literacy. After the items were compiled, they were reviewed by the same six experts. There was only one learning outcome that was identified as not measuring the intended learning outcome of statistical literacy, so a new item was developed. For the remaining items, reviewer feedback was used to refine the items to increase the validity evidence that the assessment was measuring statistical literacy.

The cognitive interviews and pilot test provided additional evidence that the assessment was measuring statistical literacy rather than measuring irrelevant content. Items were modified if students' responses did not appear to match the intended learning

outcomes. One item was deleted after the pilot test was conducted because it was clear that the item was not measuring the intended learning outcome.

Data collected in the field test was used to conduct a single-factor CFA to present evidence that the BLIS assessment measures one construct. The scree plot of eigenvalues and fit indices, including the TLI, CFI, and RMSEA, showed evidence of a single factor. Considering the results from the expert reviews, cognitive interviews, and pilot test, it can be assumed that there is validity in that the BLIS assessment measures one factor, statistical literacy.

An analysis based on IRT was conducted to see the extent to which validity exists for students of different ability levels. Using the PC model, item difficulties were computed. There are more items with low difficulty levels than items with high difficulty levels, meaning there is more evidence of validity when examining students with low ability levels compared with students with high ability levels. Results suggest that inferences made from the BLIS assessment have high validity.

Value. Evidence of the value of the BLIS assessment to statistics instructors and statistics education researchers was collected from introductory statistics instructors who participated in the field test. The value was rated on a 4-point scale from “Not at all valuable” to “Very valuable.” The value ratings for instructors and educational researchers were both high. Approximately 86% of raters indicated the assessment was valuable or very valuable, and no raters said the assessment was not at all valuable. All raters said the value to educational researchers was valuable or very valuable. Therefore,

the evidence shows that the BLIS assessment is valuable to instructors, and valuable or very valuable to educational researchers.

Reconceptualizing Statistical Literacy

As a result of conducting this study, the construct of statistical literacy was reconceptualized. The definition used during the development of the BLIS assessment was “the ability to read, understand, and communicate statistical information.” Multiple definitions of statistical literacy had been seen in the literature, such as knowing the basic language of statistics (Garfield et al., 2002) and communicating, interpreting, and being critical of statistical information (Gal, 2002). The definition used in this study was chosen partially because it relates to the definition of general literacy, the ability to read and write (“Literacy,” n.d.a; “Literacy,” n.d.b). Prior to developing the assessment, I proposed that the main components of statistical literacy in a simulation-based course included understanding terms, simulations, and inferential techniques. Also, other components could include computing descriptive statistics, understanding definitions and terms, creating graphs, and interpreting visual representations of data.

During the development of the assessment, a conception of what statistical literacy means expanded. When examining existing assessment items, the importance of using real-world contexts in the items became clearer, as recommended by Watson (2011), Gal (1998), and Garfield et al. (2005). A majority of the existing assessment items that did not include a real-world context did not appear to align with the definition of statistical literacy used to develop the assessment. Further, it has also been mentioned that one of the reasons simulation-based methods should be included in an introductory

statistics course is that creating simulations mimics a real-world problem (Cobb, 2007; Hesterberg, 1998; Rossman, 2007). The modified definition of statistical literacy revised at the end of this study is:

The ability to read, understand, and communicate statistical information. This type of statistical information that is relevant for statistical literacy (e.g., graphical representations, descriptive statistics, inferential statistics) is encountered in daily life, such as in a media article, and involves real contexts.

Limitations

The results from this study show evidence of high reliability, validity, and value of the BLIS assessment; however, there are limitations to the claims that can be made. Not all sources of measurement error that can affect reliability were examined. Participant variability can include test fatigue and lack of test taker motivation (Weathington et al., 2010), and these factors were not examined in this study. In the field test, most students received credit for completing the assessment; however, when communicating with the instructors, some students were given credit for completion, regardless of how well they did on the assessment. Receiving credit only for completion could affect students' motivation to try to do well on the assessment. Environmental variability is another source of measurement error that was not taken into account. Environmental variability results from distractions and differences in testing locations. A majority (83.3%) of students took the assessment outside of class, and this suggests that possible measurement error related to environmental issues.

Data was collected during the middle and end of the semester, and this means that validity evidence does not exist for using the BLIS assessment as a pretest. According to Thorndike and Thorndike-Christ (2010), a separate validity argument is needed for a pretest and posttest because the inferences to be made for each type of test are different. Therefore, if it is of interest to use the BLIS assessment as a pretest, data needs to be collected from students at the beginning of an introductory statistics course to collect evidence of validity.

Instructors and students who participated in the field test were not selected randomly and thus could be systematically different from the population of interest. This could affect generalizability. This is particularly a problem for the evidence of value of the BLIS assessment. The study found evidence of high value for the BLIS assessment; however, the results could be biased due to the voluntary nature of the study. It could be argued that if instructors did not find value in the assessment, they might have chosen not to participate in the field test.

Implications for Teaching

The BLIS assessment could be used by introductory statistics instructors who incorporate simulation-based methods in their course. Based on the reliability and validity evidence collected, the assessment would ideally be used towards the end of a course. The results could inform instructors about which topics their students struggle with and which topics they understand. Instructors could use the results to help their students as well as to make changes in their future courses.

Introductory statistics instructors could use the results from the BLIS assessment to guide them to teach topics that students struggled with on the assessment. There were two items that students performed poorly on in the field test, Items 21 and 35. Students chose one of the distractors in these items more frequently than students chose the correct option. For Item 21, the learning outcome was: understanding that a confidence interval for a proportion is centered at the sample statistic. Students were asked to interpret the center of the confidence interval and approximately 75% of students chose an interpretation that included being 95% confident. The learning outcome for Item 35 was: understanding of the factors that allow a sample of data to be generalized to the population. Results from Item 35 suggest that some students incorrectly believed that a random sample of size 500 was too small to make a generalization from a sample to a population when estimating a proportion.

Implications for Future Research

New items should be developed for the BLIS assessment for a couple of reasons. First, recall that one item was deleted after the pilot test because it was determined that the item was not measuring the intended learning outcome: understanding that every model is based on assumptions which limit our scope of inferences. This was a learning outcome that was suggested by one of the expert reviewers, so a new item should be written for a future version of the assessment. The second reason why new items are needed is because there are more items with lower difficulty levels than items with higher difficulty levels. More difficult items are needed to increase the precision in estimating ability levels for higher achieving students.

Student demographic information collected in the field test could be used to examine if Differential Item Functioning (DIF) exists for any of the items on the BLIS assessment. For example, items could be examined to see if males and females perform differently. Race and age categories are other variables that could be examined.

The data collected from instructors who administered the BLIS-3 assessment to their students in the field test could be used to conduct additional analyses. In particular, analysis could be conducted to examine how the extent to which simulation was incorporated in the courses was related to the performance of students on the assessment. The invitation that was sent to the instructors asked for participants who “teach an introductory statistics course at the postsecondary level that includes simulation, to some extent, in the curriculum.” However, in the survey, three instructors said they did not use bootstrap confidence intervals, randomization tests, probability simulations, or any other simulations to help understand statistical topics. The students in the courses taught by those three instructors were kept in the analyses of the field test data because when they were removed, coefficient alpha based on the individual item scores was approximately the same, .82, compared to .83 when they were included in the analysis. As mentioned previously, the average total score from the field test was 21.41 out of 36 ($s=6.25$, $N=940$). The average total score for students enrolled in a course taught by one of the three instructors who did not use simulations was much lower than the average total score for the remaining students, 15.39 ($s=4.15$, $n=77$) and 21.95 ($s=6.12$, $n=863$), respectively. Therefore, more investigation should be conducted to understand the relationship between the amount of simulation included in the curriculum and students’ responses on

the BLIS assessment. In particular, the relationship between the amount of simulation included in the curriculum could be compared with how students perform on specific items such as items involving simulation-based methods.

Multiple comparison studies could be conducted using the BLIS assessment. Effectiveness of different teaching methods and curricula could be examined. Administering the BLIS assessment at the end of a teaching experiment could be used in a comparison to see which methods best promoted students' statistical literacy. Curricula that could be compared could differ in the amount of simulation-based methods included, or could differ in the order that methods were presented. Average scores could be compared, but responses at the item-level could also be examined. When comparing different curricula, it could be possible to find no difference in average scores, but differences could exist at the item-level indicating that different curricula promote different aspects of statistical literacy.

Different populations could be examined using the BLIS assessment. Introductory statistics students could complete the assessment to provide evidence of how statistically literate students are. In addition to administering the assessment to introductory statistics students, the assessment could be administered to pre-service teachers. Conducting a study with pre-service teachers could provide information about whether or not pre-service teachers are prepared to teach students to be statistically literate.

The BLIS assessment could be administered at many time points. Students could complete the assessment multiple times during the course to see how their responses change and develop. Further, the assessment could be administered to students months

after they complete their introductory statistics course to see how much knowledge they retained.

The relationship between statistical literacy and other learning outcomes, such as statistical reasoning and statistical thinking, could be examined. How the outcomes are related could influence the order that the outcomes are taught in the classroom. Statistical reasoning and statistical thinking could be examined with different assessments and student scores could be compared with scores on the BLIS assessment. Making this comparison could inform researchers and instructors about whether or not statistical literacy, reasoning, and thinking are developed at the same time or if statistical literacy is needed before students can develop statistical reasoning and statistical thinking.

Conclusion

Evidence collected during the development of the BLIS assessment suggests that the assessment has very good psychometric properties. The assessment has high reliability, as established by the expert reviews, cognitive interviews with students, small-scale pilot test, and a large-scale field test. The degree to which validity exists for appropriate inferences to be made from the assessment is high. Reliability and validity could be slightly improved by adding a few more difficult items. The BLIS assessment was judged to be valuable or very valuable to statistics educators.

The BLIS assessment provides a valuable addition to the statistics education community. The assessment can be used by statistics education researchers who conduct studies with introductory statistics students at the postsecondary level, and are interested in statistical literacy. Statistics instructors who are interested in understanding what

statistical literacy knowledge their students have can also benefit from using the BLIS assessment.

References

- Allen, K. (2006). The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics. (Unpublished doctoral dissertation). University of Oklahoma, Norman, OK.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.
- American Statistical Association (n.d.). Statisticians in history. Retrieved from <http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=15>
- Amherst H. Wilder Foundation (n.d.). Wilder research. Retrieved from www.wilder.org/Wilder-Research/Pages/default.aspx
- Anderson, B. L., Gigerenzer, G., Parker, S., & Schulkin, J. (2014). Statistical literacy in obstetricians and gynecologists. *Journal for Healthcare Quality, 36*(1).
- Aoyama, K., & Stephens, M. (2003). Graph interpretation aspects of statistical literacy: A Japanese perspective. *Mathematics Education Research Journal, 15*(3), 207-225.
- Atkinson, M. P., Czaja, R. F., & Brewster, Z. B. (2006). Integrating sociological research into large introductory courses: Learning content and increasing quantitative literacy. *Teaching Sociology, 34*(1), 54-64.

- Beishe, R. B., Newman, F. K., Cannon, J. Duane, C., Treanor, J., Hoecke, C. V., Howe, B. J., Dubin, G. (2004). Serum antibody responses after intradermal vaccination against influenza. *New England Journal of Medicine*, 351(22), 2286-2294.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, 88(3), 588.
- Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In Ben-Zvi, D., & Garfield, J. (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Biggs, B., & Collis, F. (1982). Evaluating the quality of learning: The SOLO Taxonomy. New York, NY: Academic Press.
- Bond, T. & Fox, C. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). London: Psychology Press.
- Buckendahl, C., & Plake, B. S. (2006). Evaluating tests. In Downing, S. M., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 725-738). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Budgett, S., & Pfannkuch, M. (2007). Assessing students' statistical literacy. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/sat07/Budgett_Pfannkuch.pdf
- Callingham, R., & Watson, J. M. (2005). Measuring statistical literacy. *Journal of Applied Measurement*, 6(1), 19.

- Catalysts for Change (2013). *Statistical thinking: A simulation approach to modeling uncertainty* (2nd ed.). Minneapolis, MN: CATALYST Press.
- Chance, B. L. (1997) Experiences with authentic assessment techniques in an introductory statistics course, *Journal of Statistics Education*, 5(3) Retrieved from <http://www.amstat.org/publications/jse/v5n3/chance.html>
- 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., & Rossman, A. J. (2006). *Investigating statistical concepts, applications, and methods*. Cengage Learning.
- Chaput, B., Girard, J. C., & Henry, M. (2011). Frequentist approach: Modelling and simulation in statistics and probability teaching. In Batanero, C., Burrill, G., & Reading, C. (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 85-95). doi: 10.1007/978-94-007-1131-0_12, Springer Science+Business Media B.V.
- Chervany, N. L., Collier Jr, R. O., Fienberg, S. E., Johnson, P. E., & Neter, J. (1977). A framework for the development of measurement instruments for evaluating the introductory statistics course. *The American Statistician*, 31, 17-23.
- Chervin, R. D., Ruzicka, D. L., Vahabzadeh, A., Burns, M. C., Burns, J. W., Buchman, S. R. (2013) The face of sleepiness: Improvement in appearance after treatment of sleep apnea. *Journal of Clinical Sleep Medicine*, 9(9). doi: <http://dx.doi.org/10.5664/jcsm.2976>

- Chick, H. L., & Pierce, R. (2013). The statistical literacy needed to interpret school assessment data. *Mathematics Teacher Education and Development*. Retrieved from <http://www.merga.net.au/ojs/index.php/mted/article/view/170>
- Chick, H. L., & Pierce, R. (2012). Teaching for statistical literacy: Utilising affordances in real-world data. *International Journal of Science and Mathematics Education*, *10*(2), 339-362.
- Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1). Retrieved from: <http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art1>
- Cobb, G. W. (1998, April). The objective-format question in statistics: Dead horse, old bath water, or overlooked baby? Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Cobb, G. W. (1997). Mere literacy is not enough. In Steen, L. A. (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America* (pp. 73-90).
- Cobb, G. W. (1992). Teaching statistics. In Steen, L. (Ed.), *Heeding the call for change*, (pp. 3-43), MAA Notes No. 22, Washington: Mathematical Association of America.
- Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*, 801-823.
- College Board (2010). AP statistics course description. New York: The College Board, Retrieved from <http://apcentral.collegeboard.com/apc/public/courses/descriptions/index.html>

- Crossen, C. (1996). *Tainted truth: The manipulation of fact in America*. New York: Touchstone.
- Crowther, G. (1959). Crowther Report '15 to 18'. Report of the Central Advisory Council for Education. London, HMSO.
- delMas, R. (2013). [Homework assignment for Survey Design, Sampling, and Implementation Class]. Unpublished assignment.
- delMas, R. C. (2002). Statistical Literacy, Reasoning, and Learning: A Commentary. *Journal of Statistics Education*, 10(3) Retrieved from http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html
- delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. Paper presented at the International Research Forum on Statistical Reasoning, Thinking and Literacy Conference, Auckland, New Zealand.
- 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, Retrieved from <http://www.stat.auckland.ac.nz/serj> © International Association for Statistical Education (IASE/ISI)
- Denning, P. J. (1997). Quantitative practices. In Steen, L. A. (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America*, (pp. 106-117).
- De Veaux, R. D., & Velleman, P. F. (2008). Math is music: Statistics is literature (or, why are there no six-year-old novelists?). *Amstat News*, 54-58.

- Dossey, J. A. (1997). National indicators of quantitative literacy. In Steen, L. A. (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America* (pp. 45-59).
- Downing, S. M. (2006a). Selected-response item formats in test development. In Downing, S. M., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Downing, S. M. (2006b). Twelve steps for effective test development. In Downing, S. M., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Duggan, M. (2013). *Cell phone activities 2013*. Washington, D.C.: Pew Research Center. Retrieved from http://www.pewinternet.org/files/old-media/Files/Reports/2013/PIP_Cell%20Phone%20Activities%20May%202013.pdf
- Educational Development Center (n.d.). What is literacy? Retrieved March 9, 2013, from http://www.edc.org/newsroom/articles/what_literacy
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1-26.
- Engel, J. (2010). On teaching bootstrap confidence intervals. In *International Conference on Teaching Statistics*. Ljubljana, Slovenia: International Statistical Institute.
- Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, 19, 676-685.

- Fisher, R. A. (1936). "The Coefficient of Racial Likeness" and the Future of Craniometry. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 66, 57-63.
- François, K., Monteiro, C., & Vanhoof, S. (2008). Revealing the notion of statistical literacy within the PISA results. *status: published*.
- Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (GAISE) project: A pre-k-12 curriculum framework. Retrieved from: <http://www.amstat.org/education/gaise/>
- Gal, I. (2011). Does CensusAtSchool develop statistical literacy? *Statistical Journal of the IAOS*, 27, 229–230. doi: 10.3233/SJI-2011-0737
- Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1), 1–25.
- Gal, I. (1998). Assessing statistical knowledge as it relates to students' interpretation of data. *Reflections on statistics: Learning, teaching, and assessment in grades K-12*, 275-295.
- Gal, I. (1997). Numeracy: Imperatives of a forgotten goal. In Steen, L. A. (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America* (pp. 36-44).
- Gal, I., & Murray, S. (2011). Users' statistical literacy and information needs: Institutional and educational implications. *Statistical Journal of the IAOS*, 27(3-4), 185-195.

- Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine*, 170(5), 462-468.
- Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, 2(1), 22-38. Retrieved from [http://www.stat.auckland.ac.nz/~iase/serj/SERJ2\(1\).pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(1).pdf)
- Garfield, J., Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., & Witmer, J. (2005). Guidelines for assessment and instruction in statistics education (GAISE) project: College report. Retrieved from: <http://www.amstat.org/education/gaise/>
- 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. Doi: 10.1111/j/1751-5823.2007.00029.x
- Garfield, J., & delMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics*, 32(1).
- Garfield, J., delMas, R., & Chance, B. (2003). The web-based ARTIST: Assessment Resource Tools for Improving Statistical Thinking. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Garfield, J., delMas, R., & Chance, B. (2002). The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project. NSF CCLI grant ASA-0206571. Retrieved from <https://apps3.cehd.umn.edu/artist/index.html>

- Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinking in an introductory, tertiary-level statistics course. *ZDM – The International Journal on Mathematics Education*, 44(7), 883-898. doi: 10.1007/s11858-012-0447-5
- Garfield, J., delMas, R., & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In Bidgood, P., Hunt, N., & Jolliffe, F. (Eds.), *Assessment methods in statistical education* (pp. 75-86). John Wiley & Sons, Ltd
- Garfield, J., & Franklin, C. (2011). Assessment of Learning, for Learning, and as Learning in Statistics Education. In Batanero, C., Burrill, G., & Reading, C. (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 311-322). New York: Springer.
- Garfield, J., & Gal, I. (1999). Teaching and assessing statistical reasoning. In L. V. Stiff (Ed.), *Developing mathematical reasoning in grades K–12 (NCTM 1999 Yearbook*, pp. 207–219). Reston, VA: National Council of Teachers of Mathematics.
- Gould, R., & Ryan, C. N. (2013). *Introductory statistics: Exploring the world through data*. Pearson.
- Haack, D. G. (1979). Teaching statistical literacy. *Teaching Statistics*, 1(3), 74-76.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333.

- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450, 557-559.
- Hesterberg, T. C. (1998). Simulation and bootstrapping for teaching statistics. In *Proceedings of the Section on Statistical Education* (pp. 44-52). *The American Statistical Association*.
- Hesterberg, T., Monaghan, S., Moore, D. S., Clipson, A., & Epstein, R. (2003). Bootstrap Methods and Permutation Tests [Companion Chapter]. In Moore, D. S., McCabe, G. P., Duckworth, W. M., & Sclove, S. L. (Authors) *The Practice of Business Statistics*. New York: W. H. Freeman and Company.
- Holcomb, J., Rossman, A., & Chance, B. (2011). Exploring student understanding of significance in randomization-based courses. In *Proceedings of the 58th World Statistical Congress. Dublin, Ireland: International Statistical Institute* (pp 880-889).
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.
- Internet Legal Research Group (n.d.). *Public Legal*. Retrieved from <http://www.ilrg.com/>

- Johnson, R. W. (2001). An introduction to the bootstrap. *Teaching Statistics*, 23, 49-54.
doi: 10.1111/1467-9639.00050
- Jones, G. A., Langrall, C. W., Mooney, E. S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In Ben-Zvi, D., & Garfield, J. (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 97-117). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking, *Mathematical Thinking and Learning*, 2, 269-307.
- Jordan, E. W. (1981). Questioning strategies and sample problems for a course in statistical literacy. In *Proceedings of the Section on Statistical Education* (p. 103). *The American Statistical Association*.
- Kane, M. T. (1990). An argument based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kaplan, J. J., & Thorpe, J. (2010). Post secondary and adult statistical literacy; Assessing beyond the classroom. In *International Conference on Teaching Statistics*. Ljubljana, Slovenia: *International Statistical Institute*.
- Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). Adult literacy in America. *Washington, DC: National Center for Education Statistics*.
- Kolata, G. (1997). Understanding the news. In Steen, L. A. (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America* (pp. 23-29).

- Konold, C., & Miller, C. (2011). Tinkerplots™ Version 2 [computer software].
Emeryville, CA: Key Curriculum Press.
- Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In Garfield, J. B. & Burrill, G. (Eds.), *Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 IASE Round Table Conference*, 151-167. Voorburg, The Netherlands: International Statistical Institute.
- Lehohla, P. (2002). Promoting statistical literacy: A South African perspective. In *International Conference on Teaching Statistics*. Cape Town, South Africa: International Statistical Institute.
- Lemon, J. (2006). Plotrix: a package in the red light district of R. *R-News*, 6(4): 8-12.
- Linn, R. L. (2006). The standards for educational and psychological testing: Guidance in test development. In Downing, S. M., & a, T. M. (Eds.), *Handbook of test development* (pp. 27-38). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Literacy [Def. 1]. (n.d.a). In *Cambridge Dictionaries Online*. Retrieved March 9, 2013, from <http://dictionary.cambridge.org/dictionary/american-english/literacy?q=literacy>
- Literacy. (n.d.b). In *Merriam-Webster Online*. Retrieved March 9, 2013, from <http://www.merriam-webster.com/dictionary/literacy>
- Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the power of data*. Hoboken, NJ: Wiley.

- Lovett, M. (2001). A collaborative convergence on studying reasoning processes: A case study in statistics. In S. M. Carver & D. Klahr (Eds.), *Cognition and instructions: Twenty-five years of progress*. Hillsdale, NJ: Erlbaum, 347-384.
- Martinez-Dawson, R. (2010). The effects of a course on statistical literacy upon students' challenges to statistical claims made in the media. (Unpublished doctoral dissertation). Clemson University, Clemson, SC.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- McGrath, D. J. (2008). Comparing TIMSS with NAEP and PISA in mathematics and science. Washington, D.C.: U.S. Department of Education. Retrieved from http://nces.ed.gov/timss/pdf/Comparing_TIMSS_NAEP_%20PISA.pdf
- McGuire, C. (1993). Perspectives in assessment. *Academic Medicine*, 68(2), S3-8.
- McClave, J. T., & Sincich, T. (2007). *Statistics* (11th ed.). Upper Saddle River, NJ: Pearson.
- Mednick, S., D. Cai, J. Kanady, & S. Drummond, S. (2008). Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory. *Behavioural Brain Research*, 193. 790-86. www.elsevier.com/locate/bbr (accessed April 22, 2010).
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: American Council of Education and Macmillan.

- Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, 4, 22-63. doi: 10.1207/S15327833MTL0401_2
- Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, 65(2), 123-165. doi: 10.1111/j.1751-5823.1997.tb00390.x
- Moore, D., McCabe, G., Duckworth, W. M., & Alwan, L. (2008). *The practice of business statistics: Using data for decisions* (2nd ed.). New York, NY: W. H. Freeman.
- Moore, D. S., & Notz, W. I. (2013). *Statistics: Concepts and controversies*. New York: W. H. Freeman and Company.
- Moore, D. S., & Notz, W. I. (1979). *Statistics: Concepts and controversies*. San Francisco, CA: W. H. Freeman.
- Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology*, 2(2), 65-72.
- Muthén, B. O., DuToit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus: Statistical analysis with latent variables* (Version 6; 6th ed.). Los Angeles, CA: Muthén and Muthén.
- National Cancer Institute. (n.d.). *SEER stat fact sheets: Prostate cancer*. Retrieved January 8, 2014 from <http://seer.cancer.gov/statfacts/html/prost.html>

- National Center for Education Statistics (2011a). National assessment of educational progress (NAEP). Washington, D.C.: U.S. Department of Education, Retrieved from <http://nces.ed.gov/nationsreportcard/mathematics/>
- National Center for Education Statistics (2011b). Program for international student assessment (PISA). Washington, D.C.: U.S. Department of Education, Retrieved from <http://nces.ed.gov/surveys/pisa/>
- National Center for Education Statistics (2011c). Trends in international mathematics and science study (TIMSS). Washington, D.C.: U.S. Department of Education, Retrieved from <http://nces.ed.gov/timss/>
- Newble, D. I., Baxter, A., & Elmslie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education, 13*(4), 263-268.
- OECD (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD Survey of Adult Skills*. OECD Publishing.
doi: 10.1787/9789264128859-en
- Ohio Literacy Resource Center (2012, November 5). What is literacy? Retrieved March 9, 2013, from http://literacy.kent.edu/Oasis/Workshops/facts/whatis_lit.html
- Park, J. (2012). Developing and validating an instrument to measure college students' inferential reasoning in statistics: An argument-based approach to validation. (Doctoral Dissertation, University of Minnesota, 2012). Retrieved from <http://iase-web.org/Publications.php?p=Dissertations>
- Pew Research Center. (n.d.). *Data*. Retrieved from <http://www.pewresearch.org/data/>

- Pew Research Center. (2013). *Anonymity*. Retrieved from <http://www.pewinternet.org/datasets/july-2013-anonymity-omnibus/>
- Pew Research Center. (2011). *Reading habits*. Retrieved from <http://www.pewinternet.org/datasets/december-2011-reading-habits/>
- Pierce, R., & Chick, H. (2013). Workplace statistical literacy for teachers: Interpreting box plots. *Mathematics Education Research Journal*, 25(2), 189-205. doi: 10.1007/s13394-012-0046-3
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>
- Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York: Routledge.
- Reasoning [Def. 1]. (n.d.). In *Merriam-Webster Online*. Retrieved April 25, 2013, from <http://www.merriam-webster.com/dictionary/reasoning>
- Reed-Rhoads, T., Murphy, T. J., & Terry, R. (2006). The Statistics Concept Inventory. Retrieved from Purdue University, The Statistics Concepts Inventory project Web site: <https://engineering.purdue.edu/SCI>
- Revelle, W. (2014). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <http://CRAN.R-project.org/package=psych> Version = 1.4.5.
- Ridgway, J., Nicholson, J., & McCusker, S. (2008). Reconceptualising ‘statistics’ and ‘education’. In Batanero, C., Burrill, G., Reading, C., & Rossman, A. (Eds.),

- Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 311-322). New York: Springer.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17 (5), 1-25. Retrieved from <http://www.jstatsoft.org/v17/i05/>
- Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34(4), 441-456.
- Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *Journal of Educational Measurement*, 40(2), 163-184.
- Romeu, J. L. (1995). Simulation and statistical education. In Alexopoulos, C., Kang, K., Lilegdon, W. R., & Goldsman, D. (Eds.), *Proceedings of the 1995 Winter Simulation Conference*, 1371-1375. Arlington, VA.
- Rossmann, A. (2007). Seven Challenges for the Undergraduate Statistics Curriculum in 2007. *Slides at <http://www.statlit.org/pdf/2007RossmannUSCOTS6up.pdf>. Handout at www.statlit.org/pdf/2007RossmannUSCOTS.pdf.*
- Rossmann, A., & Chance, B. (2008). Concepts of statistical inference: A randomization-based curriculum. Retrieved from <http://statweb.calpoly.edu/csi/>
- Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, 10(3). Retrieved from <http://www.amstat.org/publications/jse/v10n3/rumsey2.html>

- Sanchez, J. (2010). The Millennium Goals, National Statistical Offices, the International Statistical Literacy Project and Statistical Literacy in Schools.
- Schiold, M. (2006). Statistical literacy survey results: Reading graphs and tables of rates and percentages. *Conference of the International Association for Social Science Information Service and Technology (IASSIST)*. Retrieved from www.StatLit.org/pdf/2006SchioldIASSIST.pdf
- Schiold, M. (2004). Statistical literacy curriculum design. *IASE Curriculum Design Roundtable*. Retrieved from www.StatLit.org/pdf/2004SchioldIASE.pdf
- Schiold, M. (1999). Statistical literacy: Thinking critically about statistics. *Of Significance*, 1(1), 15-20.
- Science World Report. (n.d.). Retrieved from www.scienceworldreport.com
- Sharma, S., Doyle, P., Shandil, V., & Talakia'atu, S. (2012). Developing statistical literacy with Year 9 students: A collaborative research project. *Proceedings of the British Society for Research into Learning Mathematics*, 32(3), 167-172.
- Shaughnessy, J. M. (2007). Research on Statistics Learning and Reasoning. In Lester, F. K. (Ed.), *Second handbook of research on mathematics teaching and learning* (2) (pp. 957-1008). Information Age Publishing Inc.
- Simon, J. L., Atkinson, D. T., & Shevokas, C. (1976). Experimental results of a radically different teaching method. *The American Mathematical Monthly*, 83(9), 733-739.
- Simon, J. L., & Bruce, P. (1990). Probability and statistics the resampling way: Stats for poets, politicians – and statisticians. Retrieved from <http://www.resample.com/content/teaching/texts/chance.txt>

- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, 28(3), 237-247.
- Smith, M. (2002). The rocky road to statistical literacy. Presentation at the ACCOLEDS conference, Vancouver, Canada.
- Snell, L. (1999a). Chance. Retrieved March 11, 2013 from <http://www.dartmouth.edu/~chance/index.html>
- Snell, L. (1999b). Using Chance media to promote statistical literacy. In *Joint Statistical Meetings, Dallas, TX*.
- Sonoda, H., Kohnoe, S., Yamazato, T., Satoh, Y., Morizono, G., Shikata, K., Morita, M., Watanabe, A., Morita, M., Kakeji, Y., Inoue, F., & Maehara, Y. (2011). Colorectal cancer screening with odour material by canine scent detection. *Gut*, 60(6), 814-819. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3095480/>. doi: 10.1136/qut.2010.218305
- S-Plus (2007). Enterprise Developer Version 8.1.1 for Microsoft Windows [computer software]. Palo Alto, CA: TIBCO Software Inc..
- Steinberg, D. M., Levine, E. L., Askew, S., Foley, P., Bennett, G. G. (2013). Daily text messaging for weight control among racial and ethnic minority women: Randomized controlled pilot study. *Journal of Medical Internet Research*, 15(11).
- Sundre, D., Thelk, A., & Wigtil, C. (2008). The Quantitative Reasoning Test, Version 9 (QR 9): Test Manual. *Harrisonburg, VA: Center for Assessment and Research Studies*. http://www.madisonassessment.com/uploads/qr-9_manual_2008.pdf

- Swanson, T., Tintle, N., VanderStoep, J., Holmes, V., & Quisenberry, B. (2010). *An active approach to statistical inference using randomization methods*. Poster presented at Joint Statistical Meetings, Vancouver, British Columbia.
- Tanis, E. A. (1992). Computer simulations to motivate understanding. In Gordon, F., & Gordon, S. (Eds.), *Statistics for the twenty-first century*, (pp. 217-225), MAA Notes No. 26. Washington: Mathematical Association of America.
- Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson.
- Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2013). Introduction to statistical investigations. Retrieved March 15, 2013, from <http://www.math.hope.edu/isi/>
- Tintle, N., Topliff, K., Vanderstoep, J., Holmes, V., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, 11(1), 21-40.
- Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, 19(1).
- Turegun, M. (2011). A model for developing and assessing community college students' conceptions of the range, interquartile range, and standard deviation. (Unpublished doctoral dissertation). University of Oklahoma, Norman, OK.
- Retrieved from <http://www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php>

- United Nations Educational, Scientific and Cultural Organization (2003, June). What is literacy? Retrieved March 9, 2013, from <http://www.unescobkk.org/resources/online-materials/aims/uis-aims-activities/uis-aims-and-literacy-assessment/>
- United States. H. R. 751-102nd Congress: National Literacy Act of 1991 (1991). In www.GovTrack.us. Retrieved April 8, 2013, from <http://www.govtrack.us/congress/bills/102/hr751>
- U.S. Census Bureau. (2014). *American Community Survey*. Retrieved April 11, 2014, from <http://www.census.gov/acs>
- Utts, J. M. (2005). *Seeing through statistics*. Belmont, CA: Duxbury Press.
- Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, 57(2), 74-79.
- Wade, B. A. (2009). Statistical literacy in adult college students. (Unpublished doctoral dissertation). The Pennsylvania State University, University Park, PA.
- Walker, H. M. (1951). Statistical literacy in the social sciences. *The American Statistician*, 5(1), 6-12.
- Wallman, K., K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, 88(421), 1-8.
- Warnes, G. R. Includes R source code and/or documentation contributed by Bolker, B., Lumley, T., & Johnson, R. C.. Contributions from Johnson, R. C. are Copyright SAIC-Frederick, Inc. Funded by the Intramural Research Program of the NIH, National Cancer Institute and Center for Cancer Research under NCI Contract

- NO1-CO-12400. (2013). gmodels: Various R programming tools for model fitting. R package version 2.15.4.1. <http://CRAN.R-project.org/package=gmodels>
- Watson, J.M. (2011). Foundations for improving statistical literacy. *Statistical Journal of the IAOS* 27, 27, 197-204.
- Watson, J. M. (2004). Developing reasoning about samples. In Ben-Zvi, D., & Garfield, J. (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277-294). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Watson, J.M. (1997). Assessing statistical thinking using the media. In I. Gal & J.B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). Amsterdam: IOS Press.
- Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3-46.
- Watson, J., & Chance, B. (2012). Building Intuitions about Statistical Inference Based on Resampling. *Australian Senior Mathematics Journal*, 26(1), 6-18.
- Watson, J. M., & Kelly, B. A. (2003). The vocabulary of statistical literacy. In *Proceedings of the joint conference of the New Zealand Association for Research in Education and the Australian Association for Research in Education, Auckland, New Zealand*. Retrieved from <http://aare.edu.au/03pap/wat03297.pdf>
- Weathington, B. L., Cunningham, C. J., & Pittenger, D. J. (2010). *Research methods for the behavioral and social sciences*. Hoboken, NJ: Wiley.
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer.

- Wild, C.J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223–265.
- Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.
- Willse, J. T. (2014). CTT: Classical Test Theory Functions. R package version 2.1. <http://CRAN.R-project.org/package=CTT>
- Wood, M. (2004). Statistical inference using bootstrap confidence intervals. *Significance*, 1, 180-182. doi: 10.1111/j.1740-9713.2004.00067.x
- World Literacy Foundation Inc. (n.d.). FAQ. *World Literacy Foundation*. Retrieved March 10, 2013, from <http://www.worldliteracyfoundation.org/faq.html>
- Yolcu, A. (2012). An investigation of eighth grade students' statistical literacy, attitudes towards statistics and their relationship. (Unpublished doctoral dissertation). Middle East Technical University, Ankara, Turkey.
- 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 manuscript. A pre-dissertation paper, University of Minnesot.
- Ziegler, L., & Garfield, J. (2013). Exploring students' intuitive ideas of randomness using an iPod shuffle activity. *Teaching Statistics*, 35(1), 2-7. doi: 10.1111_j.1467-9639.2012.005

Appendix A

Versions of the BLIS Test Blueprint

Table A1

Preliminary Test Blueprint

Topic	Learning Outcome
Samples and populations	Understanding of the difference between a sample and population
Randomness	Understanding that there are some recognizable characteristics of randomly sampled or randomly generated data Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term
Random samples	Understanding of the purpose of random sampling in an observational study Understanding that statistics computed from random samples tend to be centered at the parameter
Random assignment	Understanding of the purpose of random assignment in an experiment
Observational studies and experiments	Ability to determine what type of study was conducted
Variables	Ability to determine if a variable is quantitative or categorical Ability to determine if a variable is an explanatory variable or a response variable

(continued)

Table A1 (continued)

Topic	Learning Outcome
Statistics and parameters	Understanding of the difference between a statistic and parameter Understanding that statistics vary Understanding of resistant statistics
Equally likely outcomes	Understanding of the gambler's fallacy
Dotplots	Ability to describe and interpret a dotplot Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data Ability to create an appropriate graph to display quantitative data Understanding the importance of creating graphs prior to analyzing data
Empirical distributions	Understanding of what an empirical distribution represents Understanding that an empirical distribution shows how sample statistics tend to vary Understanding of the difference between mean of a sample, mean of a simulated sample, and mean of an empirical distribution
Bootstrap distributions	Ability to create a bootstrap distribution to estimate a proportion Ability to create a bootstrap distribution to estimate a difference in two proportions Understanding that simulated statistics in the tails of a bootstrap distribution are not plausible estimates of a population parameter Understanding that a bootstrap distribution tends to be centered at the sample statistic

(continued)

Table A1 (continued)

Topic	Learning Outcome
Randomization distributions	Ability to create a randomization distribution to test the difference between two groups
	Understanding that simulated statistics in the tails of a randomization distribution are evidence against the null hypothesis
	Understanding that a randomization distribution tends to be centered at the hypothesized null value
Proportions	Ability to interpret a probability in the context of the data
	Ability to interpret a percent in the context of the data
Mean	Ability to interpret a mean in the context of the data
	Understand how a mean is affected by skewness or outliers
Standard deviation	Ability to interpret a standard deviation in the context of the data
	Understanding of the properties of standard deviation
	Ability to estimate a standard deviation from a sample
Standard errors	Ability to estimate a standard error from an empirical distribution
	Understanding of how sample size affects the standard error
Margin of errors	Ability to interpret a margin of error
Bootstrap intervals	Understanding of the properties of a bootstrap interval
	Understanding that a bootstrap interval provides plausible values of the population parameter
	Understanding of the purpose of a bootstrap interval
	Understanding of what statistic should be computed to create a bootstrap interval

(continued)

Table A1 (continued)

Topic	Learning Outcome
Confidence levels	Understanding of how the confidence level affects the width of a bootstrap interval
Randomization tests	Understanding of the logic of a significance test when the null hypothesis is rejected Understanding of the purpose of a hypothesis test
P-values	Ability to compute a p-value using a randomization distribution
Models	Ability to describe a model for a bootstrap interval (outcomes, probabilities, with or without replacement) Ability to describe a model for a randomization test (outcomes, probabilities, with or without replacement)
Hypothesis statements	Ability to determine a null and alternative hypothesis statement based on a research question
Significance levels	Understanding of what a significance level is Understanding of how a significance level is used
Statistical significance	Ability to determine statistical significance based on a p-value
Scope of conclusions	Understanding that an experimental design with random assignment supports causal inference Understanding of the factors that allow a sample of data to be generalized to the population Understanding that correlation does not imply causation

Table A2

BLIS Test Blueprint-1

Topic	Learning Outcome and Item Number
Data production	<ol style="list-style-type: none"> 1. Understanding of the difference between a sample and population 2. Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term 3. Understanding that statistics computed from random samples tend to be centered at the parameter 4. Ability to determine what type of study was conducted 5. Ability to determine if a variable is quantitative or categorical 6. Ability to determine if a variable is an explanatory variable or a response variable 7. Understanding of the difference between a statistic and parameter 8. Understanding that statistics vary from sample to sample
Graphs	<ol style="list-style-type: none"> 9. Ability to describe and interpret a dotplot 10. 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. Understanding the importance of creating graphs prior to analyzing data
Descriptive statistics	<ol style="list-style-type: none"> 12. Ability to interpret a percent in the context of the data 13. Ability to interpret a mean in the context of the data 14. Understand how a mean is affected by skewness or outliers 15. Ability to interpret a standard deviation in the context of the data 16. Understanding of the properties of standard deviation

(continued)

Table A2 (continued)

Topic	Learning Outcome and Item Number
Empirical sampling distributions	17. Understanding of what an empirical sampling distribution represents
	18. Understanding that an empirical sampling distribution shows how sample statistics tend to vary
	19. Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter
Confidence intervals	20. Understanding that a confidence interval provides plausible values of the population parameter
	21. Understanding that a confidence interval is centered at the sample statistic
	22. Understanding of how the confidence level affects the width of a confidence interval
Randomization distributions	23. Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis
	24. Understanding of how sample size affects the standard error
	25. Understanding that a randomization distribution tends to be centered at the hypothesized null value
Hypothesis tests	26. Ability to estimate a p-value using a randomization distribution
	27. Understanding of the logic of a hypothesis test
	28. Understanding of the purpose of a hypothesis test
	29. Understanding that every model is based on assumptions which limit our scope of inferences
	30. Ability to determine a null and alternative hypothesis statement based on a research question
	31. Ability to determine statistical significance based on a p-value
	32. Understanding that errors can occur in hypothesis testing
	33. Understanding of how a significance level is used

(continued)

Table A2 (continued)

Topic	Learning Outcome and Item Number
Scope of conclusions	34. Understanding that only an experimental design with random assignment can support causal inference
	35. Understanding of the factors that allow a sample of data to be generalized to the population
Regression and correlation	36. Ability to match a scatterplot to a verbal description of a bivariate relationship
	37. Ability to use a least-squares regression equation to make a prediction

Table A3

BLIS Test Blueprint-2

Topic	Learning Outcome and Item Number
Data production	<ol style="list-style-type: none"> 1. Understanding of the difference between a sample and population 2. Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term 3. Understanding that statistics computed from random samples tend to be centered at the parameter 4. Ability to determine what type of study was conducted 5. Ability to determine if a variable is quantitative or categorical 6. Ability to determine if a variable is an explanatory variable or a response variable 7. Understanding of the difference between a statistic and parameter 8. Understanding that statistics vary from sample to sample
Graphs	<ol style="list-style-type: none"> 9. Ability to describe and interpret a dotplot 10. 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. Understanding the importance of creating graphs prior to analyzing data
Descriptive statistics	<ol style="list-style-type: none"> 12. Ability to interpret a probability in the context of the data 13. Ability to interpret a mean in the context of the data 14. Understand how a mean is affected by skewness or outliers 15. Ability to interpret a standard deviation in the context of the data 16. Understanding of the properties of standard deviation

(continued)

Table A3 (continued)

Topic	Learning Outcome and Item Number
Empirical sampling distributions	17. Understanding of what an empirical sampling distribution represents
	18. Understanding that an empirical sampling distribution shows how sample statistics tend to vary
	19. Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter
Confidence intervals	20. Understanding that a confidence interval provides plausible values of the population parameter
	21. Understanding that a confidence interval for a proportion is centered at the sample statistic
	22. Understanding of how the confidence level affects the width of a confidence interval
Randomization distributions	23. Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis
	24. Understanding of how sample size affects the standard error
	25. Understanding that a randomization distribution tends to be centered at the hypothesized null value
Hypothesis tests	26. Ability to estimate a p-value using a randomization distribution
	27. Understanding of the logic of a hypothesis test
	28. Understanding of the purpose of a hypothesis test
	29. Understanding that every model is based on assumptions which limit our scope of inferences
	30. Ability to determine a null and alternative hypothesis statement based on a research question
	31. Ability to determine statistical significance based on a p-value
	32. Understanding that errors can occur in hypothesis testing
	33. Understanding of how a significance level is used to make decisions

(continued)

Table A3 (continued)

Topic	Learning Outcome and Item Number
Scope of conclusions	34. Understanding that only an experimental design with random assignment can support causal inference
	35. Understanding of the factors that allow a sample of data to be generalized to the population
Regression and correlation	36. Ability to match a scatterplot to a verbal description of a bivariate relationship
	37. Ability to use a least-squares regression equation to make a prediction

Table A4

BLIS Test Blueprint-3

Topic	Learning Outcome and Item Number
Data production	<ol style="list-style-type: none"> 1. Understanding of the difference between a sample and population 2. Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term 3. Understanding that statistics computed from random samples tend to be centered at the parameter 4. Ability to determine what type of study was conducted 5. Ability to determine if a variable is quantitative or categorical 6. Ability to determine if a variable is an explanatory variable or a response variable 7. Understanding of the difference between a statistic and parameter 8. Understanding that statistics vary from sample to sample
Graphs	<ol style="list-style-type: none"> 9. Ability to describe and interpret a dotplot 10. 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. Understanding the importance of creating graphs prior to analyzing data
Descriptive statistics	<ol style="list-style-type: none"> 12. Ability to interpret a probability in the context of the data 13. Ability to interpret a mean in the context of the data 14. Understand how a mean is affected by skewness or outliers 15. Ability to interpret a standard deviation in the context of the data 16. Understanding of the properties of standard deviation

(continued)

Table A4 (continued)

Topic	Learning Outcome and Item Number
Empirical sampling distributions	17. Understanding of what an empirical sampling distribution represents
	18. Understanding that an empirical sampling distribution shows how sample statistics tend to vary
	19. Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter
Confidence intervals	20. Understanding that a confidence interval provides plausible values of the population parameter
	21. Understanding that a confidence interval for a proportion is centered at the sample statistic
	22. Understanding of how the confidence level affects the width of a confidence interval
Randomization distributions	23. Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis
	24. Understanding of how sample size affects the standard error
	25. Understanding that a randomization distribution tends to be centered at the hypothesized null value
Hypothesis tests	26. Ability to estimate a p-value using a randomization distribution
	27. Understanding of the logic of a hypothesis test
	28. Understanding of the purpose of a hypothesis test
	29. & 30. Ability to determine a null and alternative hypothesis statement based on a research question
	31. Ability to determine statistical significance based on a p-value
	32. Understanding that errors can occur in hypothesis testing
	33. Understanding of how a significance level is used to make decisions

(continued)

Table A4 (continued)

Topic	Learning Outcome and Item Number
Scope of conclusions	34. Understanding that only an experimental design with random assignment can support causal inference
	35. Understanding of the factors that allow a sample of data to be generalized to the population
Regression and correlation	36. Ability to match a scatterplot to a verbal description of a bivariate relationship
	37. Ability to use a least-squares regression equation to make a prediction

Appendix B

Correspondence with Expert Reviewers of the Preliminary Test Blueprint

B1 Invitation to be an Expert Reviewer

Dear Professor XXX,

I am writing to request your assistance in helping me develop the [*Basic Literacy In Statistics (BLIS) assessment*]¹ for the introductory statistics course. 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. I am working with my co-advisers, Joan Garfield and Michelle Everson.

I am requesting your help with this project because of your expertise in the areas of statistical literacy and assessment. Specifically, I would like you to review the preliminary test blueprint at this time and if you are willing, the initial [BLIS] instrument.

I define statistical literacy as being able to read, understand, and communicate statistical information. The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. This instrument includes items on modeling and simulation-based methods as well as some of the more traditional topics, because it is to be an assessment for a more modern version of the introductory course.

In light of changes to the introductory statistics course, namely the move to include modeling and simulation-based methods, a new assessment of statistical literacy is needed. A new statistical literacy assessment could be used for a variety of purposes. For example, the assessment could be used as a pretest to determine what statistical literacy skills students have prior to taking an introductory statistics course at the postsecondary level. Statistics instructors could also use the assessment in their classes to see which statistical literacy topics students understand and which topics they do not understand. Further, researchers could use the assessment to determine how much statistical literacy students gain in an introductory statistics course that incorporates modeling and simulation-based methods in the curriculum, or to determine if the amount of statistical literacy gained in introductory statistics courses taught with varying teaching methods or curricula is different.

In order to provide evidence of the validity and value of the [BLIS assessment], I am requesting your help at this time in reviewing the preliminary test blueprint.

If you agree to do this, I would like you to take the perspective of an instructor who is teaching an introductory statistics course at the postsecondary level that includes

¹ The Basic Literacy In Statistics (BLIS) assessment was originally called the Assessment of Statistical Literacy (ASL).

modeling and simulation-based methods in the curriculum. You will be asked to rate the importance of particular learning outcomes, and your feedback will be used to modify the current test blueprint and create a final version of this blueprint.

The preliminary version of the [BLIS assessment] will be created in order to measure the learning outcomes outlined in the final blueprint. I will contact you again, in [insert month], 2013, with a request for help in reviewing the preliminary version of the [BLIS assessment]. Your feedback will be invaluable as I work toward creating a final version of this assessment.

I am attaching two documents: 1) a description of statistical literacy and examples of what it means to be statistically literate and 2) the evaluation form that contains the test blueprint.

Ideally, I would like to receive a completed evaluation form from you within two weeks, by [insert date], 2013. If this will not be possible, please let me know. Further, please feel free to contact me with any questions as you evaluate the preliminary test blueprint. I sincerely hope that you will be able to help during this phase of my study, and I appreciate your time and expertise.

I hope you will agree to participate as an expert reviewer of the preliminary versions of the test blueprint and the [BLIS assessment]. Please let me know your response by replying to this email (sath0166@umn.edu).

Sincerely,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

B2 Examples of Statistical Literacy

Based on a review of the literature, I am using the following definition of statistical literacy:

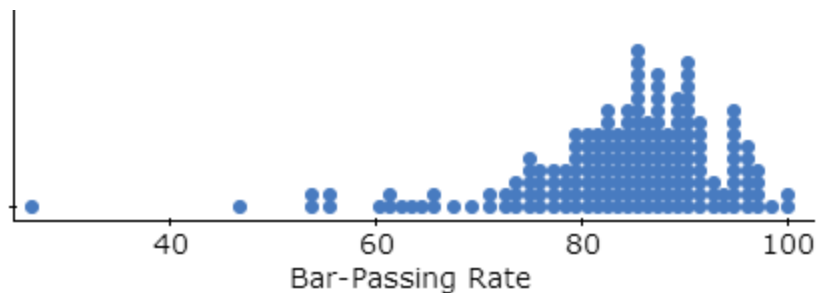
Statistical literacy is being able to read, understand, and communicate statistical information.

These three skills can be assessed by using key words such as identify, describe, translate, interpret, read, and compute.²

The examples provided below show a few skills that a statistically literate student would have after taking a modern introductory class with some mention of modeling and simulation.

Example 1 – Visualizing Bar Exam Pass Rates at Law Schools³

The Internet Legal Research Group website⁴ provides the pass rates for 185 law schools in the United States in 2006.



A statistically literate student would be able to do the following:

- Read the dotplot.
- Interpret the dotplot by making statements such as “A majority of law schools have passing rates higher than 65” and “The variability in passing rates is high, spanning from about 26% to 100%.”

² Garfield, J., & delMas, R. (2010). A web site that provides resources for assessing students’ statistical literacy, reasoning and thinking. *Teaching Statistics*, 32(1).

³ Gould, R., & Ryan, C. (2013). *Introductory statistics: Exploring the world through data*. Boston, MA: Pearson.

⁴ Internet Legal Research Group (n.d.). *Public Legal*. Retrieved from <http://www.ilrg.com/>

Example 2 – Helper or Hinderer⁵

In a study reported in the November 2007 issue of *Nature*, 10-month-old infants watched a “climber” character try to make it up a hill⁶. In one scenario, the climber was pushed to the top of the hill by a “helper” character and in the other scenario, the climber was pushed back down the hill by the “hinderer” character. After watching the two scenarios multiple times, the infant was presented with a helper and hinderer toy and 14 of the 16 infants chose the helper toy over the hinderer toy. The research question is “Are infants able to notice and react to helpful or hindering behavior observed in others?”

A statistically literate student would be able to do the following:

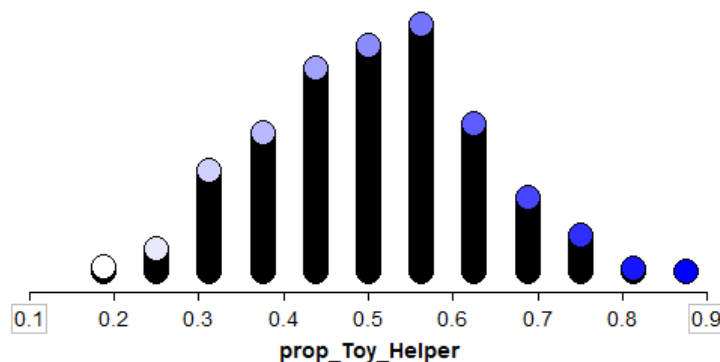
- Identify that the statistic of interest is the proportion of infants in the sample that chose the hinderer toy.
- Compute the proportion of infants in the sample that chose the hinderer toy which is $14/16 = 0.875$.

Suppose for the moment that the researchers’ conjecture is wrong, and infants *do not* really show any preference for either type of toy. In other words, infants just randomly pick one toy or the other, without any regard for whether it was the helper toy or the hinderer. This is the model based on random chance—the ‘just by chance’ model.

A statistically literate student would be able to do the following:

- Describe the random chance model as consisting of two outcomes (the helper and hinderer toys) each with a .5 probability of being chosen and the toy is chosen with replacement.

A simulation was run with 1000 trials to simulate the random chance model. A plot of the results from the simulation is shown below.



A statistically literate student would be able to do the following:

- Compute the approximate p-value based on the simulation provided above which is .001.

⁵ Catalysts for Change (2012). *Statistical thinking: A simulation approach to modeling uncertainty*. Minneapolis, MN: CATALYST Press.

⁶ J. K. Hamlin, K. Wynn, & P. Bloom. (2007). Social evaluation by preverbal infants. *Nature*, 450, 557-559.

B3 Preliminary Test Blueprint Review Form

Statistical Literacy Assessment Blueprint Review Form

The following table contains a list of potential statistical literacy topics and learning outcomes for students enrolled in an introductory statistics course at the postsecondary level. You will see that this list includes topics that involve modeling and simulation-based methods because more and more of these topics are being included in the introductory statistics curriculum.

The list of topics and learning outcomes was compiled from introductory statistics textbooks that include modeling and simulation-based methods.

Your responses to the blue print will help in two ways:

- Identify the most important topics to be on the statistical literacy assessment and
- validate the topics and learning outcomes on the blueprint.

Please note that there are currently too many topics and learning outcomes listed to measure in one assessment and your expert opinion will be used to help narrow down the topics and learning outcomes to the most important ones for a new assessment.

In your review, please take the perspective of an instructor who is teaching an introductory statistics course at the postsecondary level that includes modeling and simulation-based methods (to some extent) in the curriculum.

1. Rate how important each learning outcome listed below is in determining how statistically literate a student is by placing an X in the appropriate box.

Topic	Learning Outcome	Not essential for test 1	2	3	Essential for test 4
Samples and populations	Understanding of the difference between a sample and population				
Randomness	Understanding that there are some recognizable characteristics of randomly sampled or randomly generated data				

Topic	Learning Outcome	Not essential for test 1	2	3	Essential for test 4
Randomness	Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term				
Random samples	Understanding of the purpose of random sampling in an observational study				
	Understanding that statistics computed from random samples tend to be centered at the parameter				
Random assignment	Understanding of the purpose of random assignment in an experiment				
Observational studies and experiments	Ability to determine what type of study was conducted				
Variables	Ability to determine if a variable is quantitative or categorical				
	Ability to determine if a variable is an explanatory variable or a response variable				
Statistics and parameters	Understanding of the difference between a statistic and parameter				
	Understanding that statistics vary				
	Understanding of resistant statistics				
Equally likely outcomes	Understanding of the gambler's fallacy				
Dotplots	Ability to describe and interpret a dotplot				
	Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data				
	Ability to create an appropriate graph to display quantitative data				
	Understanding the importance of creating graphs prior to analyzing data				

Topic	Learning Outcome	Not essential for test 1	2	3	Essential for test 4
Empirical distributions	Understanding of what an empirical distribution represents				
	Understanding that an empirical distribution shows how sample statistics tend to vary				
	Understanding of the difference between mean of a sample, mean of a simulated sample, and mean of an empirical distribution				
Bootstrap distributions	Ability to create a bootstrap distribution to estimate a proportion				
	Ability to create a bootstrap distribution to estimate a difference in two proportions				
	Understanding that simulated statistics in the tails of a bootstrap distribution are not plausible estimates of a population parameter				
	Understanding that a bootstrap distribution tends to be centered at the sample statistic				
Randomization distributions	Ability to create a randomization distribution to test the difference between two groups				
	Understanding that simulated statistics in the tails of a randomization distribution are evidence against the null hypothesis				
	Understanding that a randomization distribution tends to be centered at the hypothesized null value				
Proportions	Ability to interpret a probability in the context of the data				
	Ability to interpret a percent in the context of the data				
Mean	Ability to interpret a mean in the context of the data				
	Understand how a mean is affected by skewness or outliers				
Standard deviation	Ability to interpret a standard deviation in the context of the data				

Topic	Learning Outcome	Not essential for test 1	2	3	Essential for test 4
Standard deviation	Understanding of the properties of standard deviation				
	Ability to estimate a standard deviation from a sample				
Standard errors	Ability to estimate a standard error from an empirical distribution				
	Understanding of how sample size affects the standard error				
Margin of errors	Ability to interpret a margin of error				
Bootstrap intervals	Understanding of the properties of a bootstrap interval				
	Understanding that a bootstrap interval provides plausible values of the population parameter				
	Understanding of the purpose of a bootstrap interval				
	Understanding of what statistic should be computed to create a bootstrap interval				
Confidence levels	Understanding of how the confidence level affects the width of a bootstrap interval				
Randomization tests	Understanding of the logic of a significance test when the null hypothesis is rejected				
	Understanding of the purpose of a hypothesis test				
P-values	Ability to compute a p-value using a randomization distribution				
Models	Ability to describe a model for a bootstrap interval (outcomes, probabilities, with or without replacement)				
	Ability to describe a model for a randomization test (outcomes, probabilities, with or without replacement)				
Hypothesis statements	Ability to determine a null and alternative hypothesis statement based on a research question				
Significance levels	Understanding of what a significance level is				
	Understanding of how a significance level is used				

Topic	Learning Outcome	Not essential for test 1	2	3	Essential for test 4
Statistical significance	Ability to determine statistical significance based on a p-value				
Scope of conclusions	Understanding that an experimental design with random assignment supports causal inference				
	Understanding of the factors that allow a sample of data to be generalized to the population				
	Understanding that correlation does not imply causation				

2. Please comment on any topics or learning outcomes included in the table that were not clearly described.

3. Please describe any important statistical literacy topics that were not included in the table above.

4. Please describe any important statistical literacy learning outcomes that were not included in the table above.

5. Please add any suggestions you have for improving the test blueprint.

Thank you for helping develop this test blueprint for a new Statistical Literacy assessment.

Appendix C

Versions of the BLIS Assessment

C1 Preliminary BLIS Assessment

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

Assessment and Item #: New Item – Real-world context based on pewinternet.org/

1. The Pew Research Center surveyed a nationally representative sample of 1,002 adults in 2013. The sample percent of internet users that have had an email or social networking account compromised was 21%. Identify the sample and population you would like to make inferences about.

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

Assessment and Item #: New Item

2. Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 10 times and the other student flips a coin 50 times. Which student is more likely to get close to half of their coin flips heads up? Explain why you chose the student you did.

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

Assessment and Item #: New Item

3. A manufacturer of frozen pizzas produces hamburger pizzas where the true average weight is 400 grams. To check the quality of the manufacturing process, random samples of 25 pizzas are taken daily and the average weight is recorded. Assuming that nothing is wrong with the manufacturing process, which sequence below is the most plausible for the average weight for five samples?
 - a. 381, 389, 405, 424, 441.
 - b. 336, 362, 377, 387, 400.
 - c. 395, 402, 420, 445, 450.
 - d. Any of the above.

Learning Outcome: Ability to determine what type of study was conducted
Assessment and Item #: ARTIST Topic Scale Test – Data Collection 7 – Modified

4. Suppose a researcher wanted to determine if aspirin reduces the chance of a heart attack. The researcher studied 500 patients that 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?
- Observational
 - Experimental
 - Survey
 - None of the above

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 = subcompact, 2 = compact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.

Learning Outcome: Ability to determine if a variable is quantitative or categorical
Assessment and Item #: ARTIST Topic Scale Test – Data Collection 2

5. What type of variable is this?
- categorical
 - quantitative
 - continuous

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

Assessment and Item #: ARTIST Topic Scale Test – Data Collection 3

6. The student plans to see if there is a relationship between the number of speeding tickets a student gets in a year and the type of vehicle he or she drives. Identify the response variable in this study.
- college students
 - type of car
 - number of speeding tickets
 - average number of speeding tickets last year

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

Assessment and Item #: ARTIST Website Item ID = Q0618

7. CNN conducted a quick vote poll on September 19, 1999 to determine “What proportion of Americans still think that the Miss America pageant is still relevant today?” The poll was conducted on the internet. Here are the results of the poll: Is the Miss American pageant still relevant today? Yes: 1192 votes, No: 4389 votes; Total: 5581 votes. Describe the parameter of interest.

Learning Outcome: Understanding that statistics vary from sample to sample

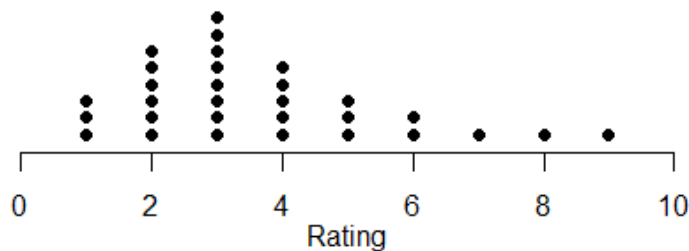
Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)

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. Researcher B claims that the first study must have been flawed because the mean was not the same in both studies. How would you respond to Researcher B’s statement?

Learning Outcome: Ability to describe and interpret a dotplot

Assessment and Item #: ARTIST Topic Scale Test – Data Representation 12 – Modified

9. One of the items on the student survey for an introductory statistics course was "Rate your aptitude to succeed in this class on a scale of 1 to 10" where 1 = Lowest Aptitude and 10 = Highest Aptitude. The instructor examined the data for men and women separately. Below is the distribution of this variable for the 30 women in the class.

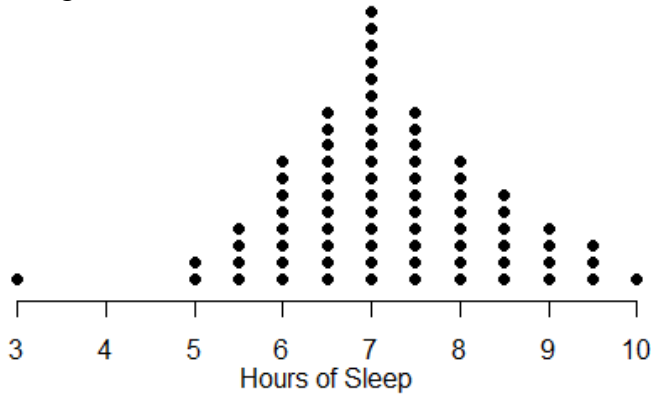


How should the instructor interpret the women's perceptions regarding their success in the class?

- A majority of women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
- The women in the class see themselves as having lower aptitude for statistics than the men in the class.
- If you remove the three women with the highest ratings, then the result will show an approximately normal distribution.

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
Assessment and Item #: CAOS 1 – Modified

10. The following graph shows a distribution of hours slept last 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.

- The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five.
- The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
- Most students seem to be getting enough 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.
- The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours.

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

Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)

11. A researcher was interested in determining if a smaller dose of a particular flu vaccine was just as successful as a full dose. A randomized experiment was conducted with 100 participants. Half of the participants received the full dose of the vaccine and the other half received a half dose of the vaccine. The number of days the participant got the flu 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 get the flu for the full dose group and half dose group. What step should the research take after the data is collected but before the hypothesis test is conducted?

Learning Outcome: Ability to interpret a percent in the context of the data
Assessment and Item #: ARTIST Topic Scale Test – Probability 9

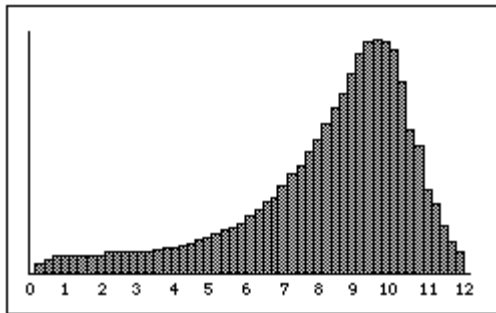
12. The local Meteorological claims that there is a 70% probability of rain tomorrow. Provide the best interpretation of this statement.
- Approximately 70% of the city will receive rain within the next 24 hours.
 - Historical records show that it has rained on 70% of previous occasions with the same weather conditions.
 - If we were to repeatedly monitor the weather tomorrow, 70% of the time it will be raining.
 - Over the next ten days, it should rain on seven of them.

Learning Outcome: Ability to interpret a mean in the context of the data
Assessment and Item #: New Item

13. According to a pet store owner, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean.

Learning Outcome: Understand how a mean is affected by skewness or outliers
Assessment and Item #: ARTIST Topic Scale Test – Sampling Variability 7

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?

- 4 to 6
- 7 to 9
- 10 to 12

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

Assessment and Item #: ARTIST Topic Scale Test – Measures of Spread 2

15. The 30 introductory statistics students took another quiz worth 30 points. On this quiz, the standard deviation of the scores of that quiz was 1 point. Which of the following gives the most suitable interpretation?
- all of the individual scores are one point apart
 - the difference between the highest and lowest score is 1
 - the difference between the upper and lower quartile is 1
 - a typical score is within 1 point of the mean

Learning Outcome: Understanding of the properties of standard deviation

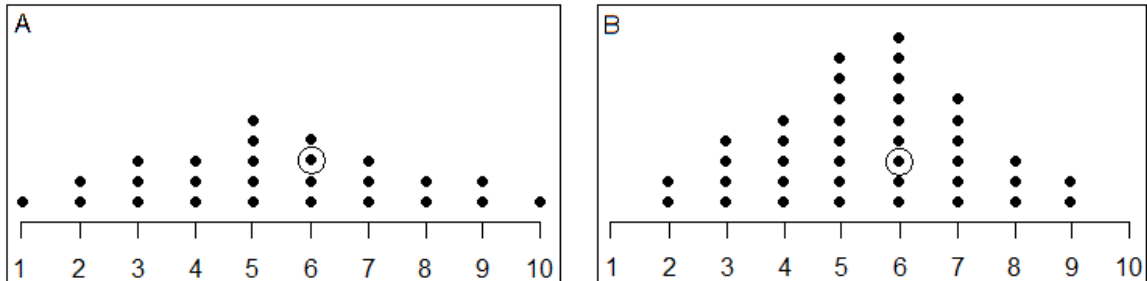
Assessment and Item #: ARTIST Topic Scale Test – Measures of Spread 5

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 for the class to be -2.30. What do we know?
- The standard deviation was calculated incorrectly.
 - Most students received negative scores.
 - Most students scored below the mean.
 - None of the above.

Learning Outcome: Understanding of what an empirical sampling distribution represents

Assessment and Item #: ARTIST Topic Scale Test – Sampling Variability 1 – Modified

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



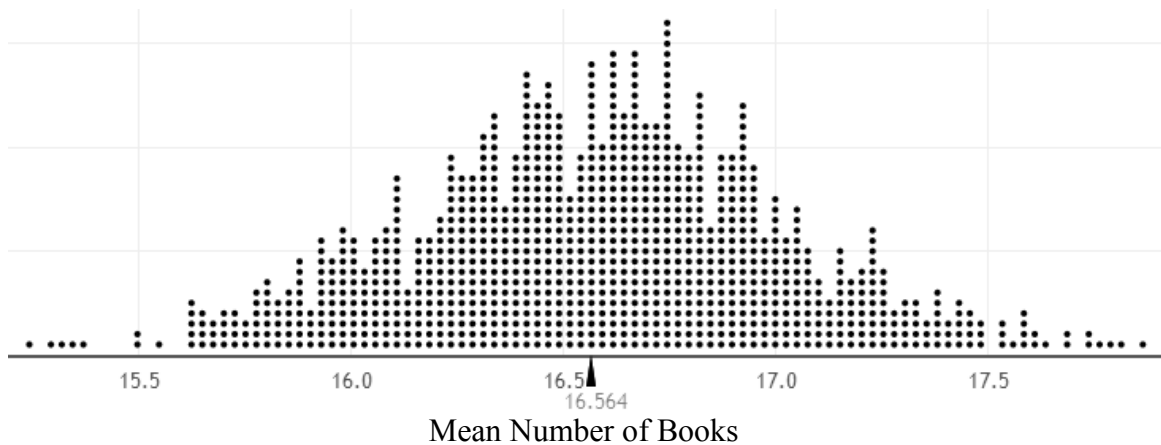
- a. No, in both Figure A and Figure B, the X represents one pebble that weights 6 grams.
- b. Yes, Figure A has a larger range of values than Figure B.
- c. Yes, the X in Figure A is the weight for a single pebble, while the X in Figure B represents the average weight of 3 pebbles.

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



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

Assessment and Item #: New Item – Real-world context based on pewinternet.org/

18. What information is obtained from this distribution?

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

Assessment and Item #: New Item – Real-world context based on pewinternet.org/

19. What values do you believe would NOT be plausible estimates of the population average number of books read? Explain your answer.

Learning Outcome: Understanding that a bootstrap interval provides plausible values of the population parameter

Assessment and Item #: New Item – Real-world context based on pewinternet.org/

20. The Pew Research Center surveyed 2,076 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% bootstrap interval was 58% to 62%. What is this interval attempting to estimate?

Learning Outcome: Understanding that a confidence interval is centered at the sample statistic

Assessment and Item #: ARTIST TestBank Item ID = Q0943 – Modified

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%)?
- We can say that 37% of veterans in the sample have been divorced at least once
 - We can say that 37% of veterans in the population have been divorced at least once
 - We can say with 95% confidence that 37% of veterans in the sample have been divorced at least once
 - We can say with 95% confidence that 37% of veterans in the population have been divorced at least once

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

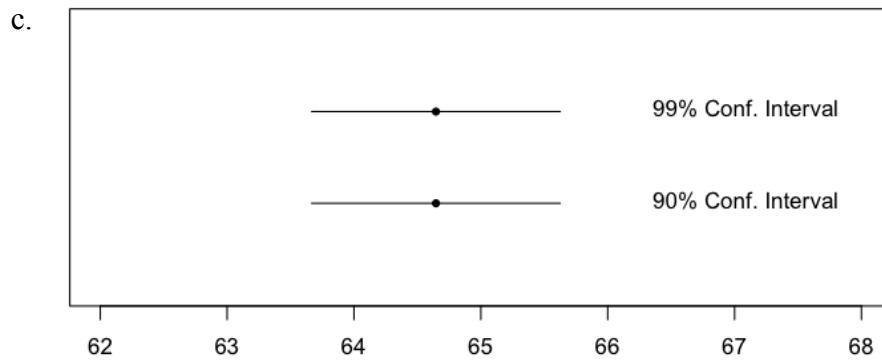
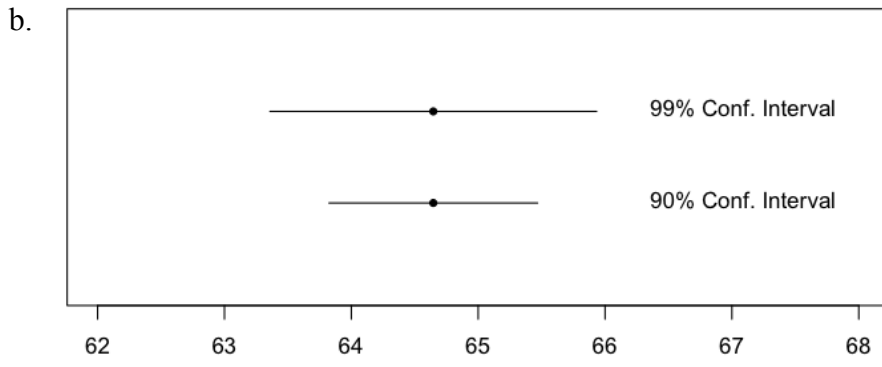
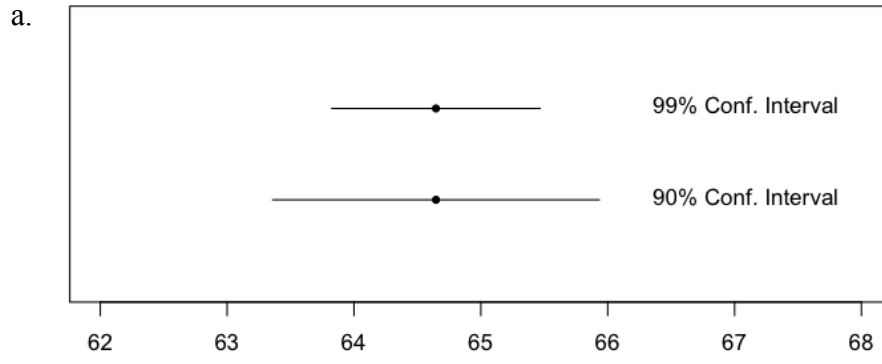
Assessment and Item #: GOALS 14 – Modified

22. This question 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.

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.

Which of the following options would best represent how the two confidence intervals would compare to each other?



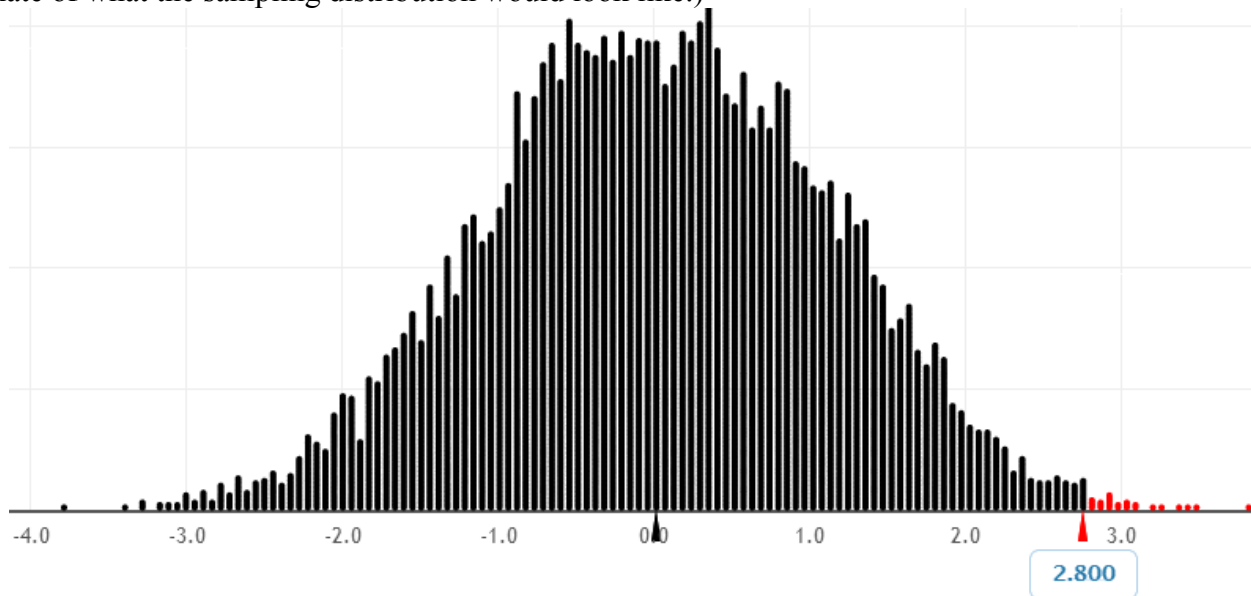
Items 23 and 24 refer to the following situation:

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.

A randomization distribution was graphed by doing the following:

- From the original sample, the 24 participants were re-randomized to the nap group (n=12) and caffeine group (n=12), without replacement.
- The mean difference in words recalled was computed for the re-randomized groups 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.)



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

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

Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)

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 for the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis? Explain your answer.

Learning Outcome: Understanding of how sample size affects the standard error
Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)

24. Suppose the sample size was doubled from 24 participants to 48 participants and the participants were still split equally into the two groups. Would the standard error change? If so, how? Explain your answer.

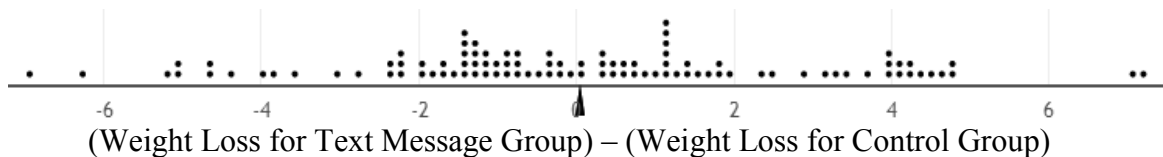
Items 25 and 26 refer to the following situation:

An experiment was conducted with 50 obese women. All women participated in a weight loss program. 26 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. The average weight loss for the text message group was 2.8 pounds and -2.6 pounds for the control group. Note that the control group had a negative weight loss which means that they actually gained weight. The difference in average weight loss was $(2.8) - (-2.6) = 5.4$.

A randomization distribution was graphed by doing the following:

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



Learning Outcome: Understanding that a randomization distribution tends to be centered at the hypothesized null value

Assessment and Item #: New Item – Real-world context based on Steinberg, Levine, Askew, Foley, and Bennett (2013)

25. Why is the randomization distribution centered at 0?

Learning Outcome: Ability to estimate a p-value using a randomization distribution
Assessment and Item #: New Item – Real-world context based on Steinberg, Levine, Askew, Foley, and Bennett (2013)

26. The alternative hypothesis is that the text messages lead to a higher weight loss than no text messages for women participating in this weight loss program. Therefore a one-tailed (i.e., one-sided) test will be conducted. Compute the p -value for the observed difference in mean weight loss of 5.4 based on the simulated data. Show how to find this p value and explain each step.

Learning Outcome: Understanding of the logic of a hypothesis test
Assessment and Item #: CAOS 40

27. The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?
- The circuit is definitely not good and needs to be repaired.
 - The electrician decides that the circuit is defective, but it could be good.
 - The circuit is definitely good and does not need to be repaired.
 - The circuit is most likely good, but it could be defective.

Learning Outcome: Understanding of the purpose of a hypothesis test
Assessment and Item #: New Item – Real-world context based on Chervin et al. (2013)

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)? 20 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. 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 statistics (70%) as evidence of the effectiveness? Explain your answer.

Learning Outcome: Understanding that every model is based on assumptions which limit our scope of inferences

Assessment and Item #: New Item

29. A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions as a model to make inferences about the two states? Explain your answer.

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

Assessment and Item #: New Item – Real-world context based on www.wilderresearch.org/

30. 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 days and 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 is difference for males and females with regards to the average number of nights spent in a place not intended for housing?” In order to conduct a hypothesis test to answer this research question, what would the null and alternative hypothesis statements be?

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

Assessment and Item #: CAOS 19

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

Learning Outcome: Understanding that errors can occur in hypothesis testing
Assessment and Item #: New Item – Real-world context based on example from Utts (2003)

32. A clinical trial was conducted to determine if women who have regular mammograms to screen for breast cancer would describe 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 as 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? Explain your answer.

Learning Outcome: Understanding of how a significance level is used
Assessment and Item #: New Item – Real-world context based on example from Lock, Lock, Lock, Lock, and Lock (2013)

33. 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 research question is “Does the dog correctly identify cancer more than half of the time?” The p-value is less than .001. Using a significance level of .05, what conclusion should be made? Explain why you chose to make your conclusion.

Learning Outcome: Understanding that only an experimental design with random assignment can support causal inference
Assessment and Item #: ARTIST Topic Scale Test – Data Collection 4

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?
- Correlational study
 - Randomized experiment
 - Time Series study
 - Survey

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

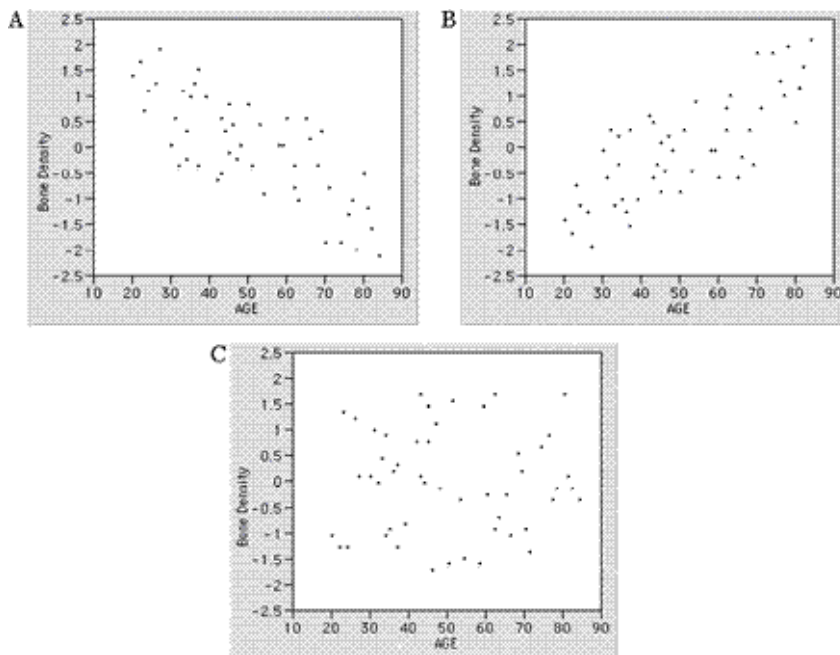
Assessment and Item #: CAOS 38

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. Which of the following does NOT affect the college official's ability to generalize the survey results to all dormitory students?
- a. Five thousand students live in dormitories on campus. A random sample of 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.

Learning Outcome: Ability to match a scatterplot to a verbal description of a bivariate relationship

Assessment and Item #: GOALS 7

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



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

Learning Outcome: Ability to use a least-squares regression equation to make a prediction

Assessment and Item #: ARTIST Topic Scale Test – Bivariate Data, Quantitative 13

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 model:

$$\text{Predicted Price} = 5620 - 440 * \text{Age}$$

A friend asked him to predict the price of a 5 year old model of this car, using his equation. Which of the following is the most correct response to provide?

- a. Plot a regression line, find 5 on the horizontal axis, and read off the corresponding value on the y axis.
- b. Substitute 5 in the equation and solve for "price".
- c. Both of these methods are correct.
- d. Neither of these methods is correct.

C2 BLIS-1 Assessment

1. The Pew Research Center surveyed a nationally representative sample 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.

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? Explain why you chose the student you did.

3. A manufacturer of frozen pizzas produces sausage pizzas, which have a true 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 pizzas in the sample is recorded. Assuming that nothing is wrong with the manufacturing process, which sequence below is the most plausible for the average weights of the five samples?
 - a. 380, 385, 413, 424, 437 (mean = 407.8, sd = 24.67)
 - b. 336, 362, 377, 387, 400 (mean = 372.4, sd = 24.64)
 - c. 396, 400, 426, 445, 449 (mean = 423.2, sd = 24.63)
 - d. Any of the above.

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
 - d. None of the above

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 = subcompact, 2 = compact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.

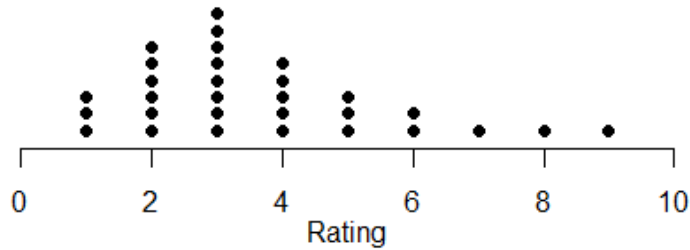
5. What type of variable is this?
 - a. categorical
 - b. quantitative
 - c. continuous

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

7. CNN conducted a quick vote poll with a random sample of 5581 Americans on September 19, 1999 to determine “What proportion of Americans still think that the Miss America pageant is still relevant today?” Here are the results of the poll: Is the Miss American pageant still relevant today? Yes: 1192 votes, No: 4389 votes. Identify the statistic and parameter of interest.

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?

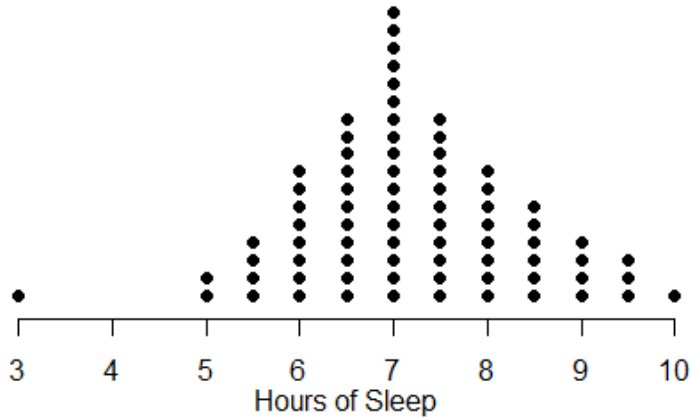
9. One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. The instructor examined the data for men and women separately. Below is the distribution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



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

- A majority of women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
- The women in the class see themselves as having lower confidence of being able to succeed in statistics than the men in the class.
- If you remove the three women with the highest ratings, then the result will show an approximately normal distribution.

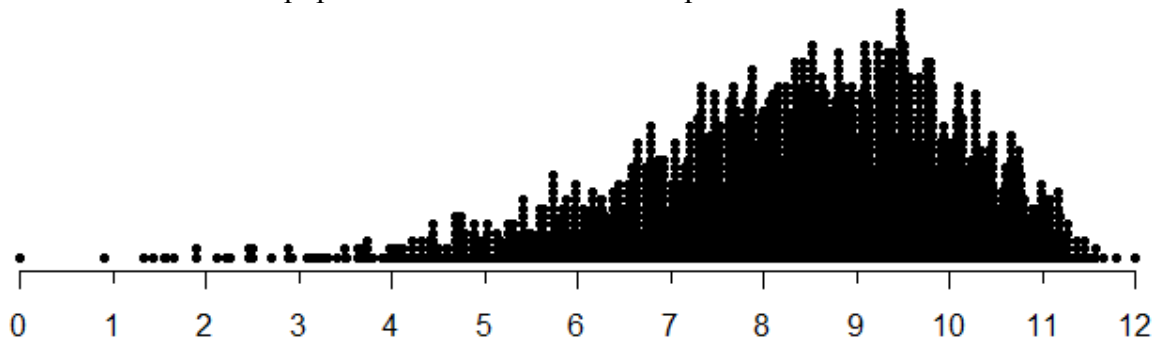
10. The following graph shows a distribution of hours slept the previous night by a group of college students.



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.

- The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five.
 - The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
 - Most students seem to be getting enough 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.
 - The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours.
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. Why should the researcher create and examine graphs of the number of days participants had flu symptoms before the hypothesis test is conducted?

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. Explain what the statistic, .15, means in the context of this report from the National Cancer Institute.
13. According to a national survey of pet owners, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean in context.
14. The distribution for a population of measurements is presented below.



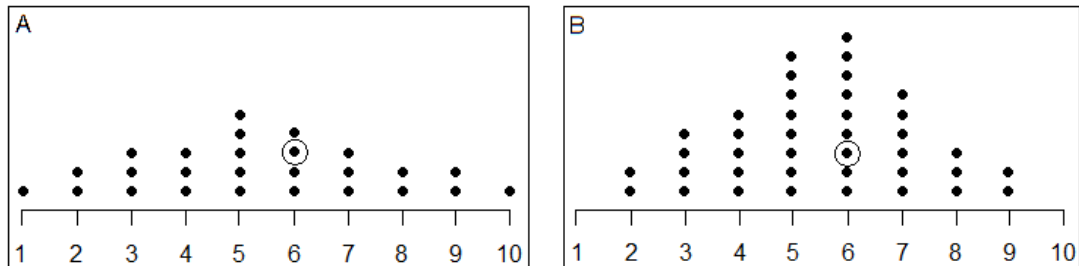
A sample of 10 randomly selected values will be taken from the population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?

- a. 6 to 7
 - b. 8 to 9
 - c. 9 to 10
 - d. 10 to 11
15. Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quick 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
 - c. the difference between the upper and lower quartile is 1
 - d. a typical distance of a score from the mean is 1 point

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.

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.



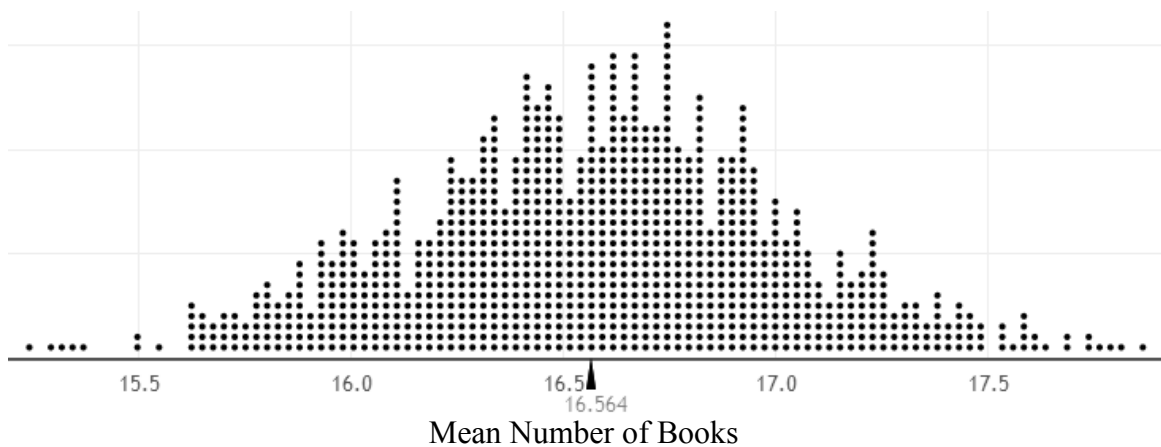
- a. No, in both Figure A and Figure B, the circled dot represents one pebble that weights 6 grams.
- b. Yes, Figure A has a larger range of values than Figure B.
- c. Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B represents the average weight of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

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.)



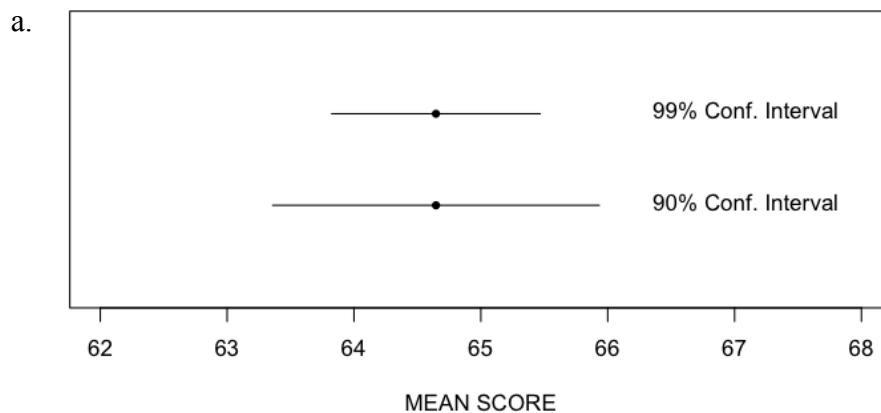
18. What information can be obtained from this distribution?
19. What values do you believe would be LESS plausible estimates of the population average number of books read if you wanted to estimate the population average with 95% confidence? Explain your answer.
20. The Pew Research Center surveyed 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?

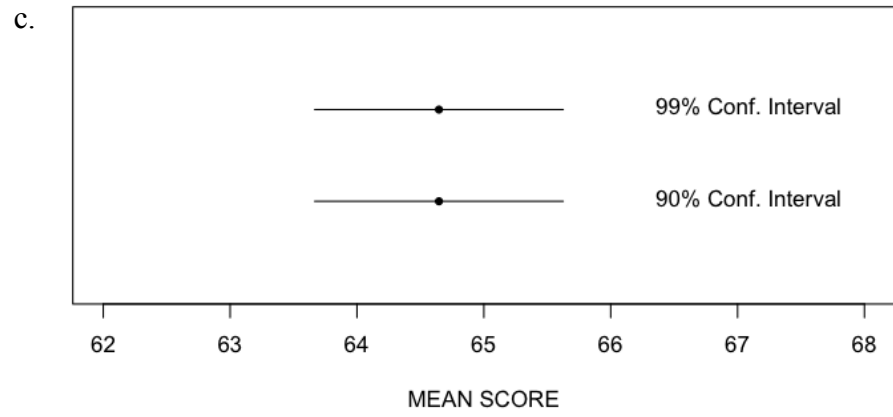
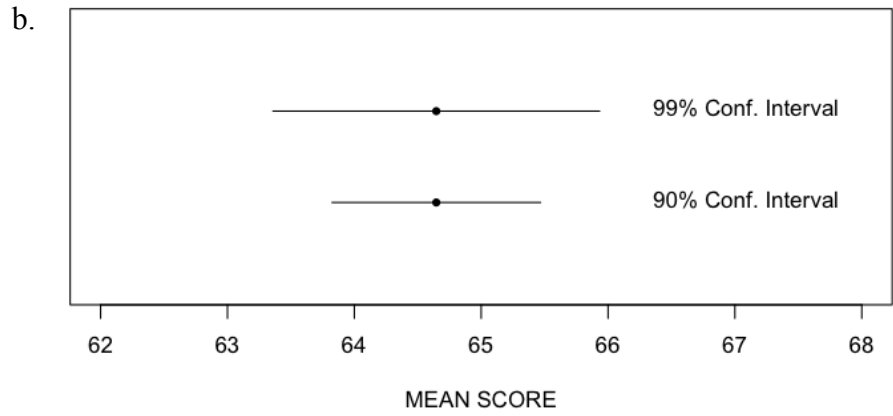
21. In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements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

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.

Which of the following options would best represent how the two confidence intervals would compare to each other?





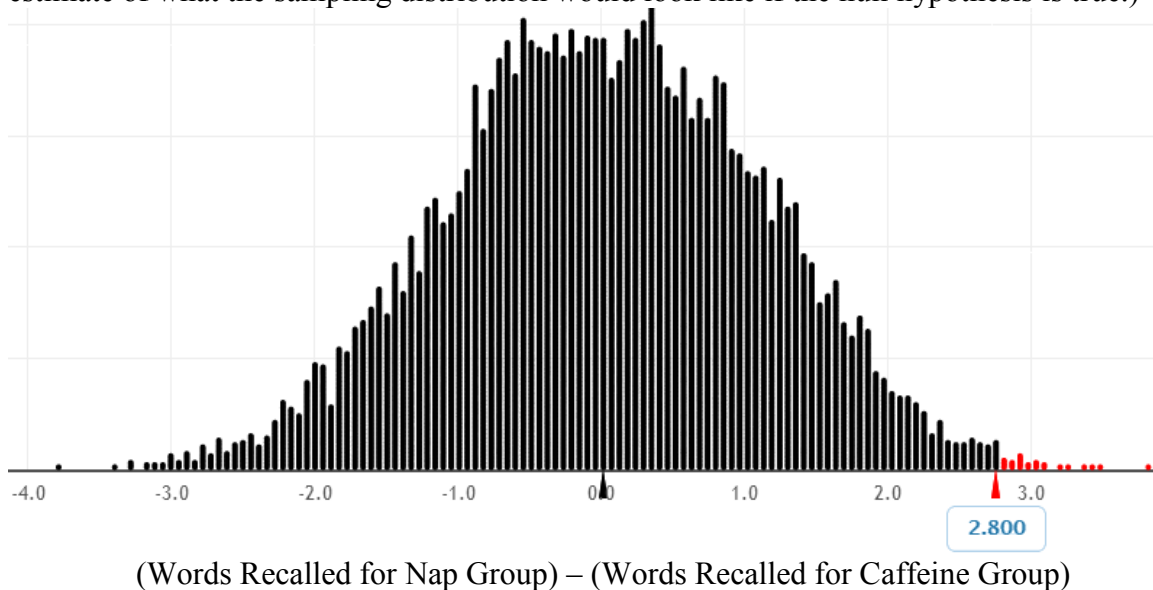
Items 23 and 24 refer to the following situation:

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.

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 [$\text{mean}(\text{nap group}) - \text{mean}(\text{caffeine group})$] 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.)



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 for the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis? Explain your answer.
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 to change? Explain your answer.

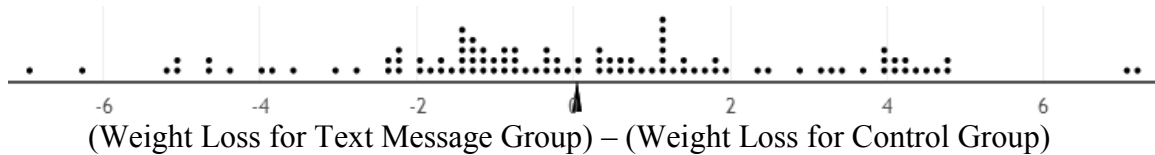
Items 25 and 26 refer to the following situation:

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 weight loss which means that they actually gained weight. The difference in average weight loss was $(2.8) - (-2.6) = 5.4$.

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 [$\text{mean}(\text{text message}) - \text{mean}(\text{control})$] and placed on the plot shown below.
- This was repeated 99 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.)



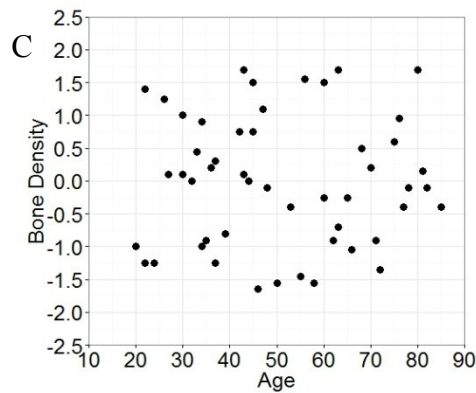
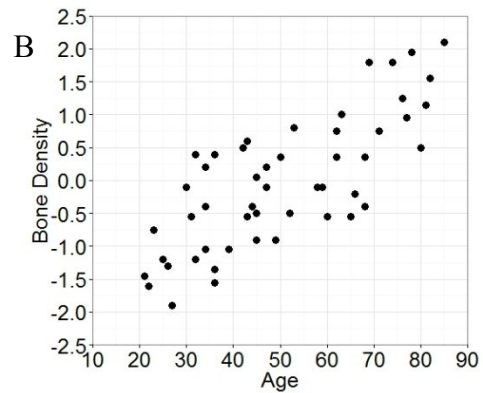
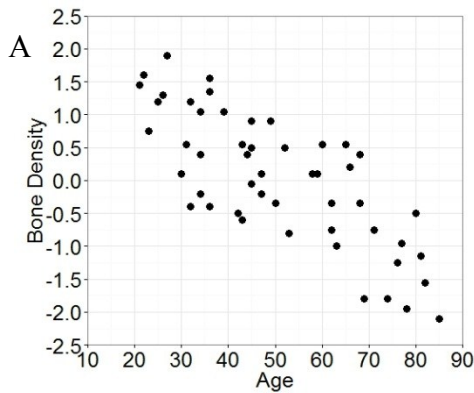
25. Why is the randomization distribution centered at 0?

26. Researchers are interested in whether 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' interest. Explain so someone else can replicate your work how you found his p -value.

27. The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?
- The circuit is definitely not good and needs to be repaired.
 - The electrician decides that the circuit is defective, but it could be good.
 - The circuit is definitely good and does not need to be repaired.
 - The electrician decides that the circuit is good, but it could be defective.
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. 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 statistics (70%) as evidence of the effectiveness? Explain your answer.
29. A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions to make inferences about the two states? Explain your answer.
30. 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 days and 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 is difference for males and females with regards to the average number of nights spent in a place not intended for housing?” In order to conduct a hypothesis test to answer this research question, what would the null and alternative hypothesis statements be?

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 large p -value.
 - A small p -value.
 - The magnitude of a p -value has no impact on statistical significance.
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 as 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? Explain your answer.
33. 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 research question is “Does the dog correctly identify cancer more than half of the time?” The p -value is less than .001. Using a significance level of .05, what conclusion should be made? Explain why you chose to make your conclusion.
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?
- Observational study
 - Randomized experiment
 - Survey

35. A college official conducted a survey 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. Which of the following does NOT affect the college official's ability to generalize the survey results to all dormitory students?
- Although 5,000 students live in dormitories on campus-only 500 were sent the survey.
 - The survey was sent to only first-year students.
 - Of the 500 students who were sent the survey, only 160 responded.
 - All of the above present a problem for generalizing the results.
36. Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?



- Graph A
- Graph B
- Graph C

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 model:

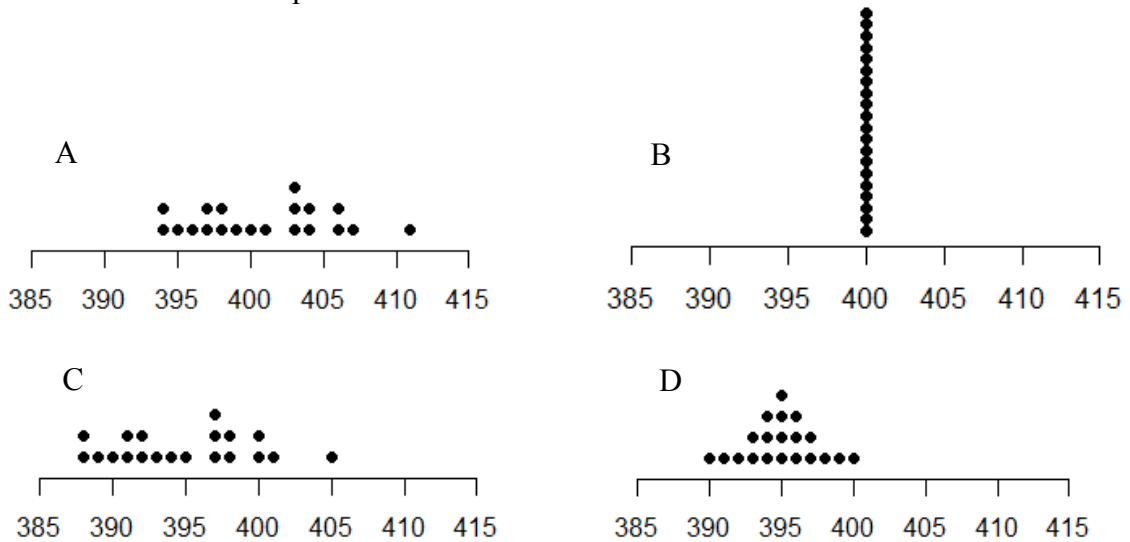
$$\text{Predicted Price} = 5620 - 440 * \text{Age}$$

A friend asked him to use his equation 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?

- Plot a regression line, find 5 on the horizontal axis, and read off the corresponding value on the y axis.
- Substitute 5 in the equation and solve for "Predicted Price".
- Both of these methods are correct.
- Neither of these methods is correct.

C3 BLIS-2 Assessment

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.
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? Explain why you chose the student you did.
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. Graph A
- b. Graph B
- c. Graph C
- d. Graph D

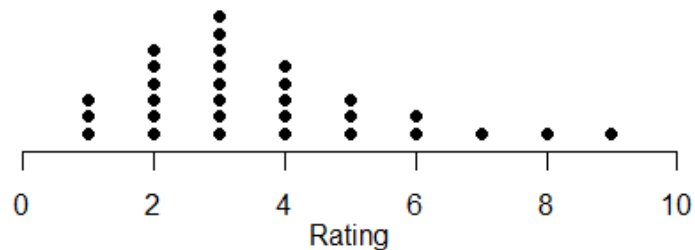
4. Suppose a researcher wanted to determine if aspirin reduces the chance of a heart attack. The researcher studied 500 patients who visited a regional hospital in the last year. Half (250) of the patients were randomly assigned to take aspirin every day and the other half to take a placebo everyday. 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

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.

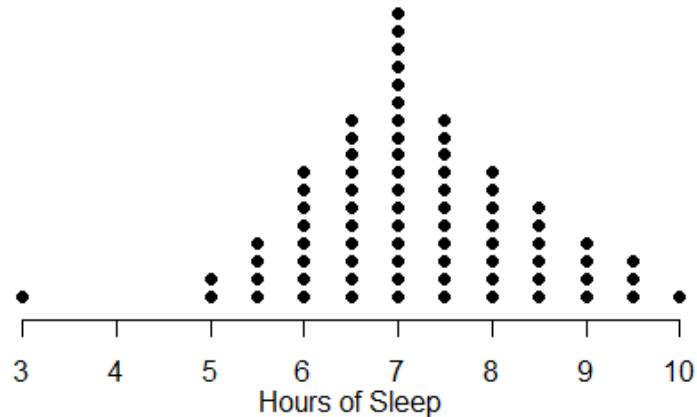
5. What type of variable is this?
 - a. Categorical
 - b. Quantitative
 - c. Continuous
6. The student plans to see if the type of vehicle a student drives is a predictor of the number of speeding tickets he or she gets in a year. Identify the response variable in this study.
 - a. College students
 - b. Type of vehicle
 - c. Number of speeding tickets
 - d. Average number of speeding tickets last year
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.

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?
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?

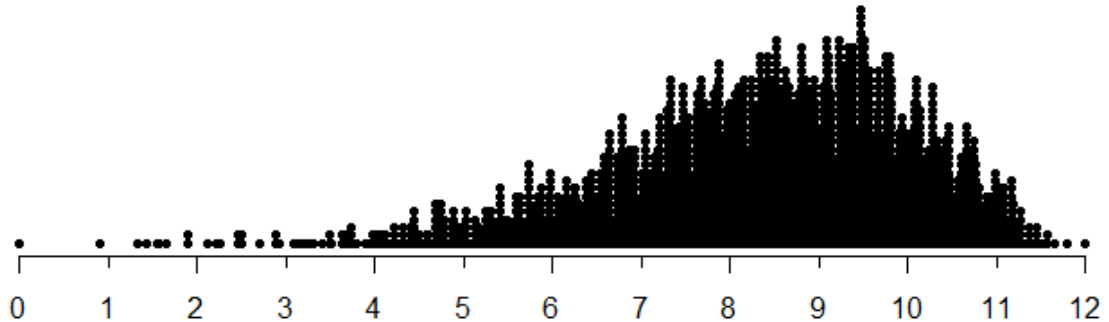
10. The following graph shows a distribution of hours slept the previous night by a group of college students.



- 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.
- 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 distribution is normal, with a mean of about 7 and a standard deviation of about 1.
 - 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.
 - 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.
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. Why should the researcher create and examine graphs of the number of days participants had flu symptoms before the hypothesis test is conducted?
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. Explain what the statistic, .15, means in the context of this report from the National Cancer Institute.

13. According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean in context.

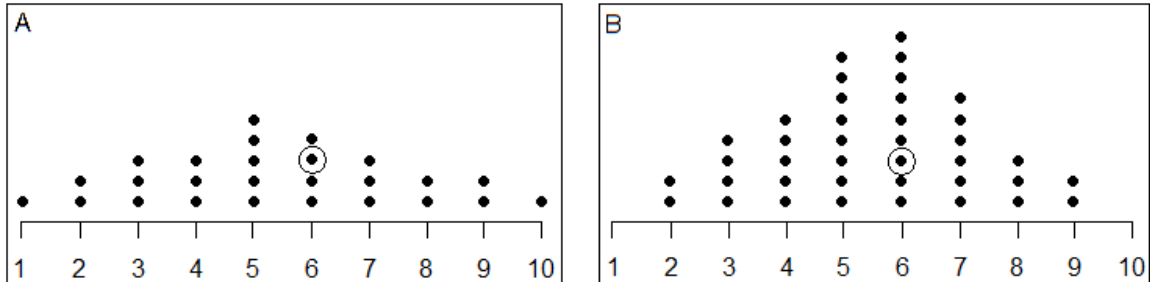
14. The distribution for a population of measurements is presented below.



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
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.
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.

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? Explain your answer.

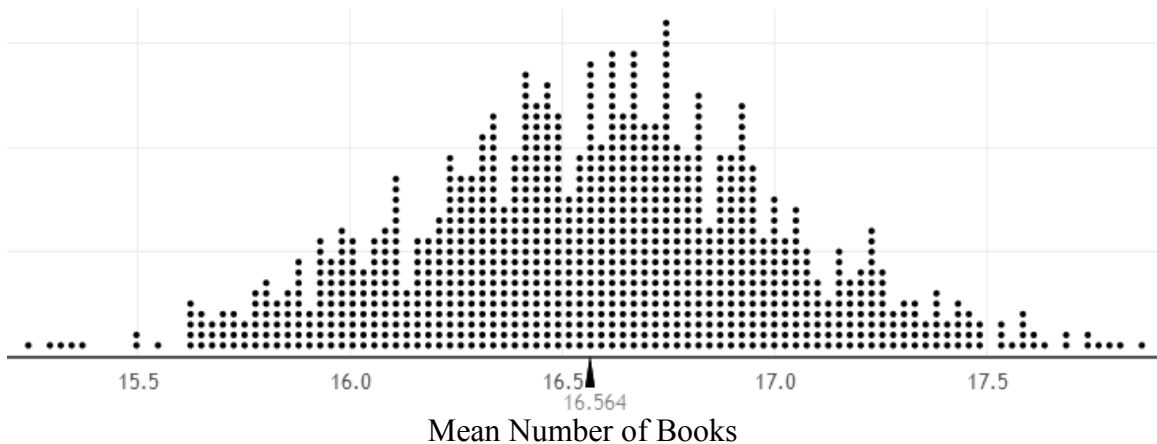


Items 18 and 19 refer to the following situation:

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.)



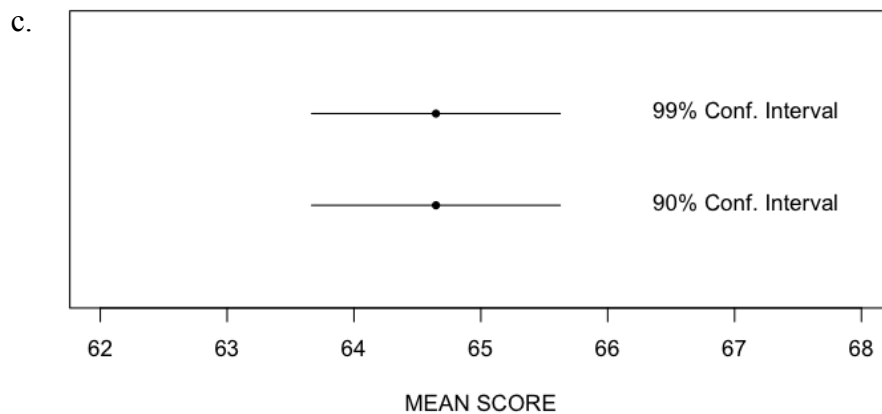
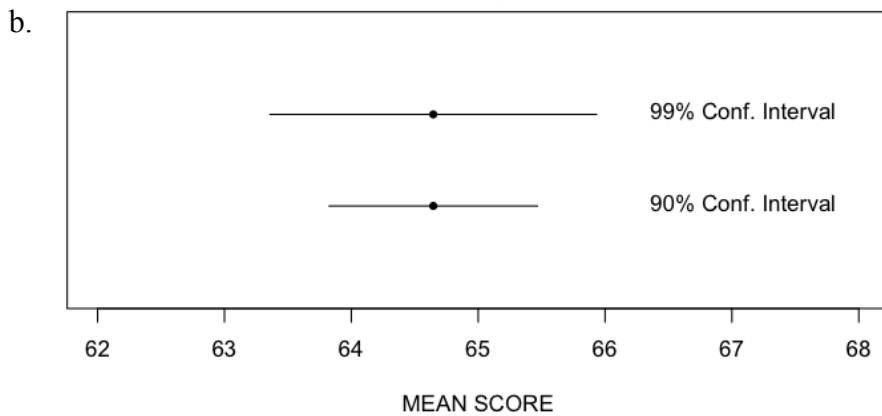
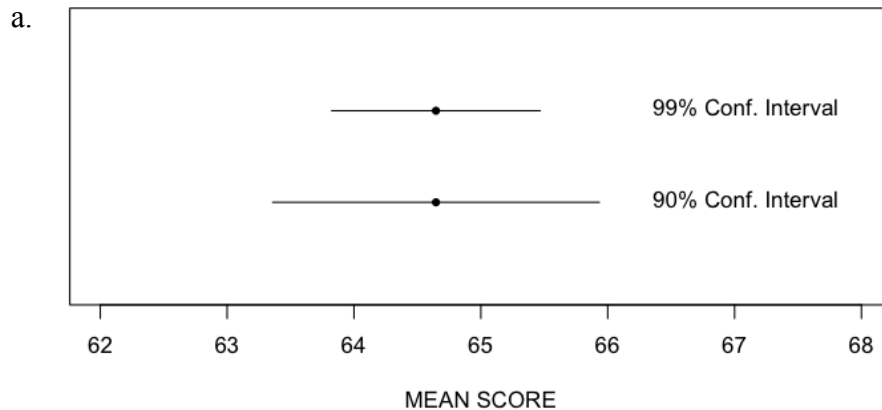
18. What information about the variability from sample to sample can be obtained from this distribution?

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? Explain your answer.
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?
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%)?
- We know that 37% of veterans in the *sample* have been divorced at least once.
 - We know that 37% of veterans in the *population* have been divorced at least once.
 - We can say with 95% confidence that 37% of veterans in the *sample* have been divorced at least once.
 - We can say with 95% confidence that 37% of veterans in the *population* have been divorced at least once.

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?



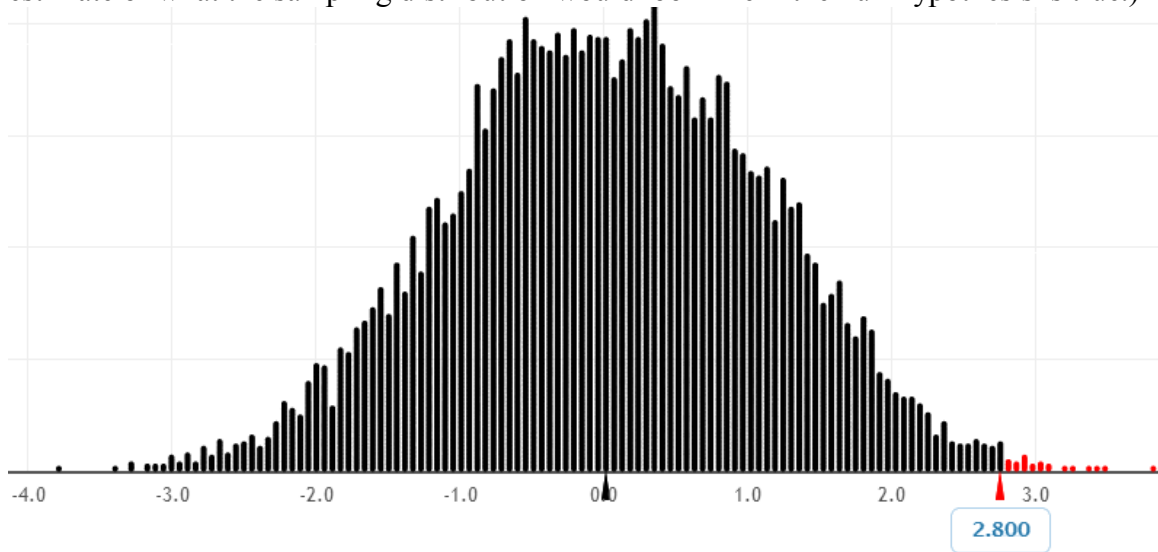
Items 23 and 24 refer to the following situation:

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 [$\text{mean}(\text{nap group}) - \text{mean}(\text{caffeine group})$] 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)

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? Explain your answer.

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? Explain your answer.

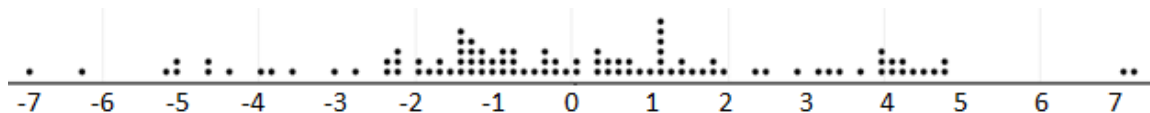
Items 25 and 26 refer to the following situation:

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. (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)

25. Explain why the randomization distribution centered at 0.
26. Researchers are interested in whether 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' interest. Explain how you found his p -value so someone else can replicate your work.

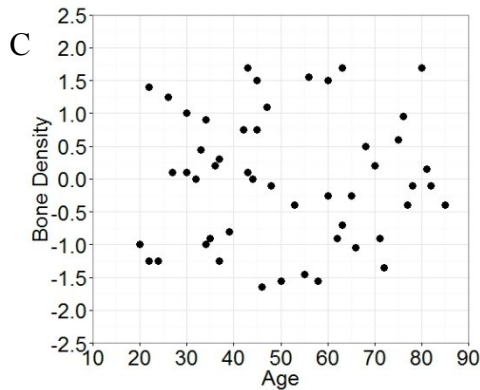
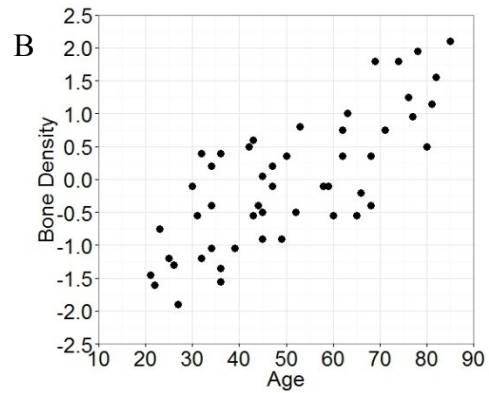
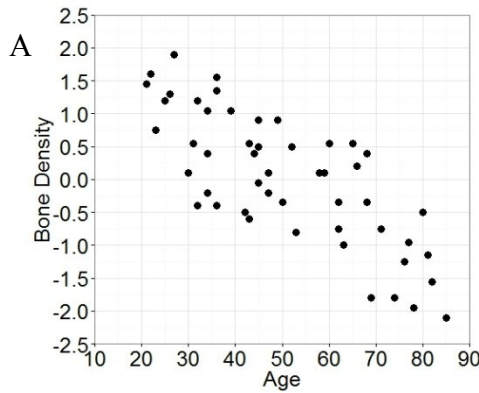
27. The following situation models the logic of a hypothesis test. An electrician tests whether or not an electrical circuit is good. The null hypothesis is that the circuit is good. The alternative hypothesis is that the circuit is not good. The electrician performs the test and decides to reject the null hypothesis. Which of the following statements is true?
- The circuit is definitely not good and needs to be repaired.
 - The circuit is most likely not good, but it could be good.
 - The circuit is definitely good and does not need to be repaired.
 - The circuit is most likely good, but it might not be good.
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. 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? Explain your answer.
29. A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions to make inferences about the two states? Explain your answer.
30. 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?” In order to conduct a hypothesis test to answer this research question, what would the null and alternative hypothesis statements be?

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 large p -value.
 - A small p -value.
 - The magnitude of a p -value has no impact on statistical significance.
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? Explain your answer.
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 half 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. Explain why you chose to make your decision.
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?
- Observational study
 - Randomized experiment
 - Survey

35. A college official conducted a survey 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. Which of the following does **NOT** affect the college official's ability to generalize the survey results to all dormitory students?

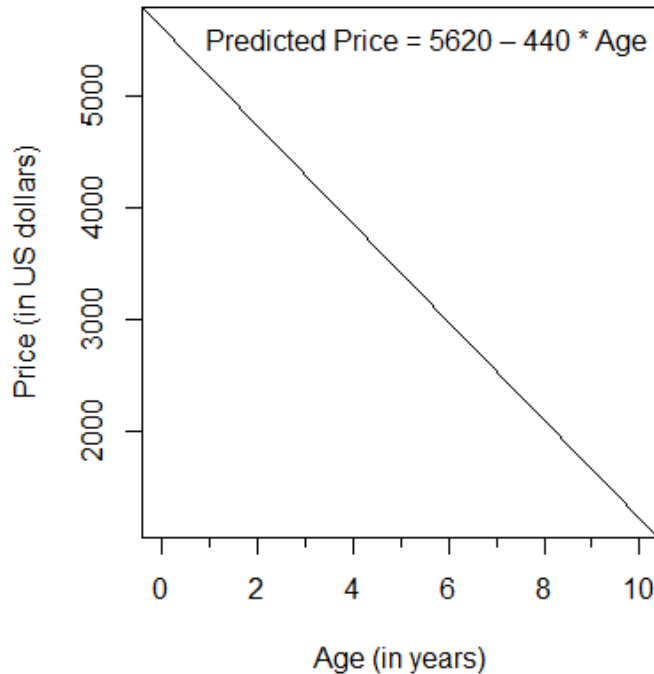
- a. Although 5,000 students live in dormitories on campus, only 500 were sent the survey.
- b. The survey was sent to only first-year students.
- c. Of the 500 students who were sent the survey, only 160 responded.
- d. All of the above present a problem for generalizing the results.

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



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

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?

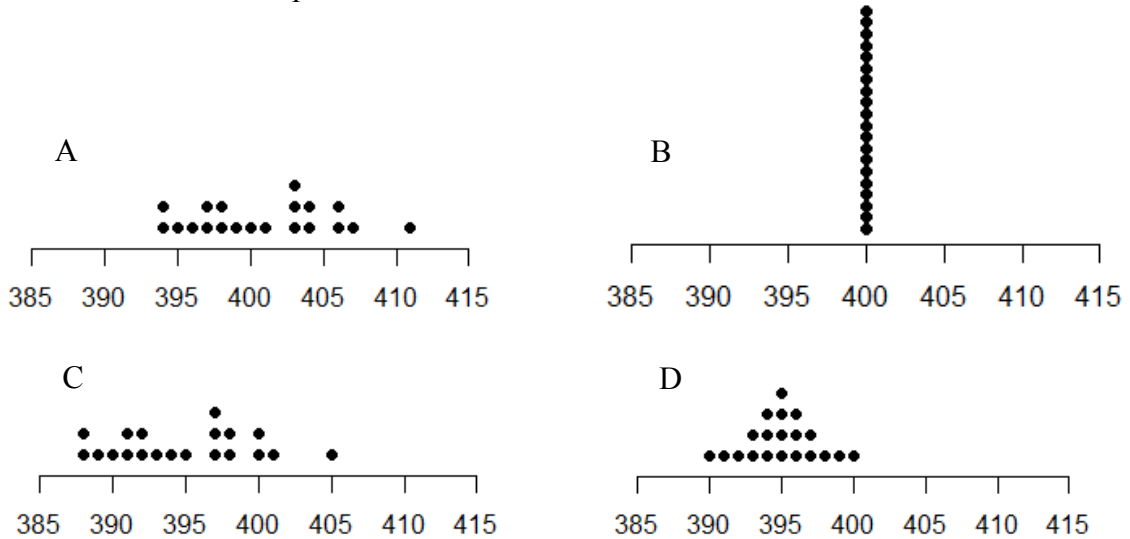
- Locate the point on the line that corresponds to an age of 5 and read off the corresponding value on the y axis.
- Substitute an age of 5 in the equation and solve for "Predicted Price".
- Both of these methods are correct.
- Neither of these methods is correct.

C4 BLIS-3 Assessment

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.

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.

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?



- Graph A
 - Graph B
 - Graph C
 - Graph D
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?
- Observational
 - Experimental
 - Survey

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.

5. What type of variable is this?
 - a. Categorical
 - b. Quantitative
 - c. Continuous

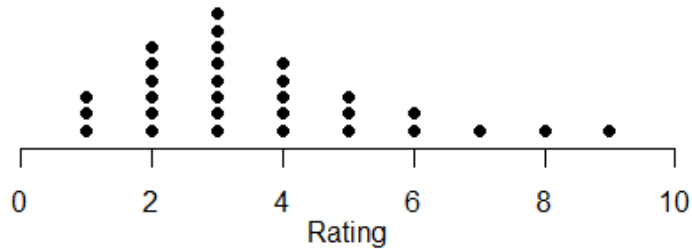
6. The student plans to see if the type of vehicle a student drives is a predictor of the number of speeding tickets he or she gets in a year. Identify the response variable in this study.
 - a. College students
 - b. Type of vehicle
 - c. Number of speeding tickets
 - d. Average number of speeding tickets last year

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.

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.

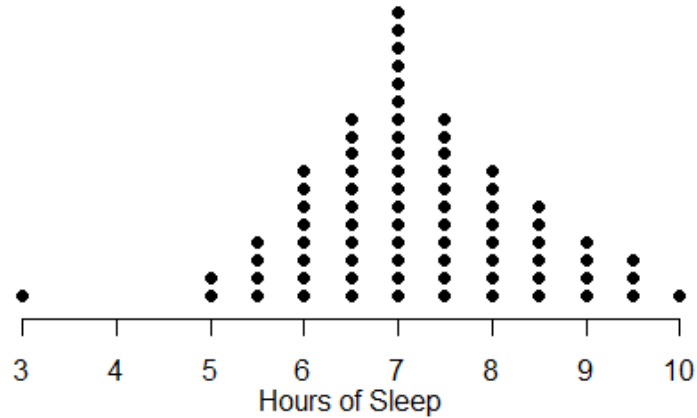
9. One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. Below is the distribution of this variable for the 30 students in the class.



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

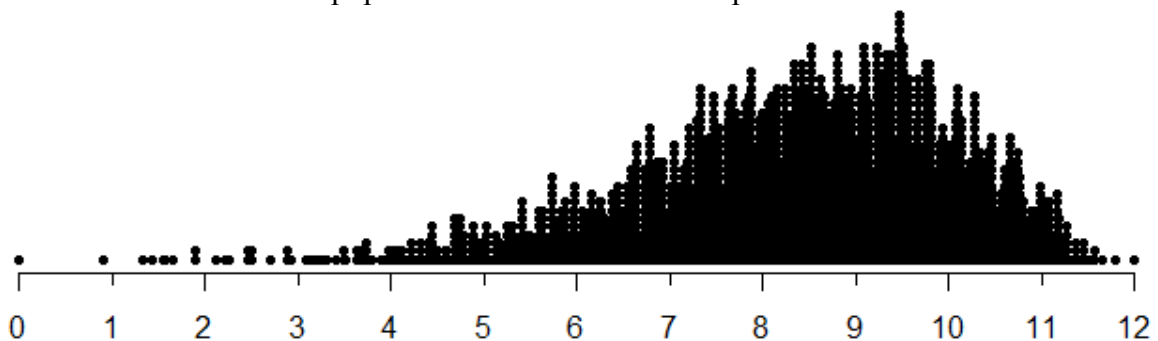
- a. A majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
- b. A majority of students in the class rated their confidence as a 3 although some ratings were higher and some ratings were lower.
- c. A majority of students will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

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.
- 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 distribution is normal, with a mean of about 7 and a standard deviation of about 1.
 - 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.
 - 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.
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?
- To decide what the null hypothesis and alternative hypothesis should be.
 - To compute the average number of days participants had flu symptoms in order to conduct a hypothesis test.
 - To see if there are recognizable differences in the two groups to decide if a hypothesis test is necessary.

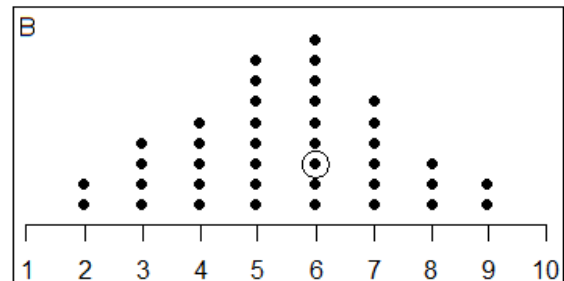
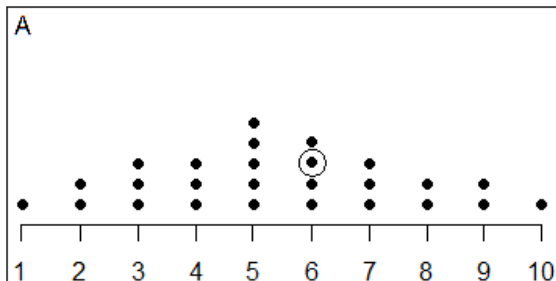
12. According to the National Cancer Institute, the probability of a man in the United States developing prostate cancer at some point during his lifetime is .15. What does the statistic, .15, mean in the context of this report from the National Cancer Institute?
- For all men living in the United States, approximately 15% will develop prostate cancer at some point in their lives.
 - 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.
 - In a random sample of 100 men in the United States, 15 men will develop prostate cancer.
 - Both a and b are correct.
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?
- 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.
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?

- 6 to 7
- 8 to 9
- 9 to 10
- 10 to 11

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?
- All of the individual scores are one point apart.
 - The difference between the highest and lowest score is 1 point.
 - The difference between the upper and lower quartile is 1 point.
 - A typical distance of a score from the mean is 1 point.
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?
- The standard deviation was calculated incorrectly.
 - Most students received negative scores.
 - Most students scored below the mean.
 - None of the above.
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.



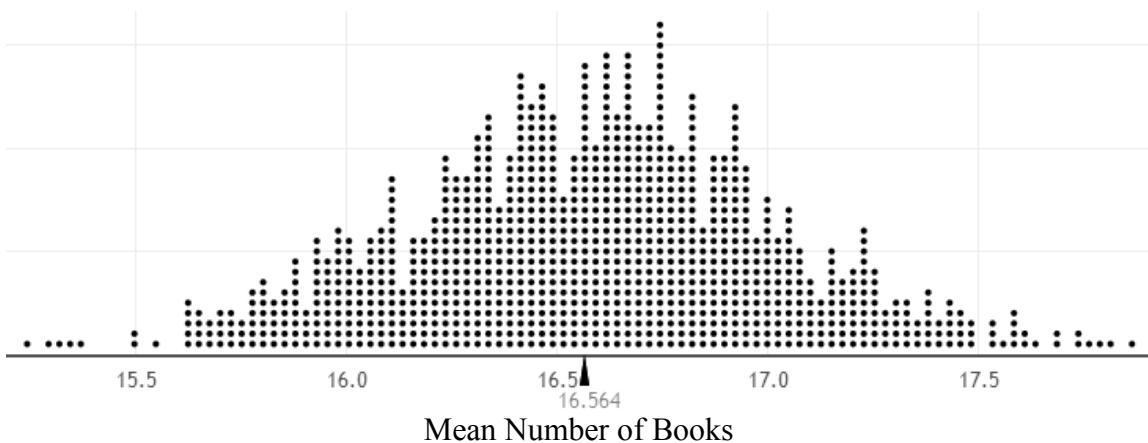
- No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
- Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6.
- Yes, the circled dot in Figure A is the weight for a single pebble, while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

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.)



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

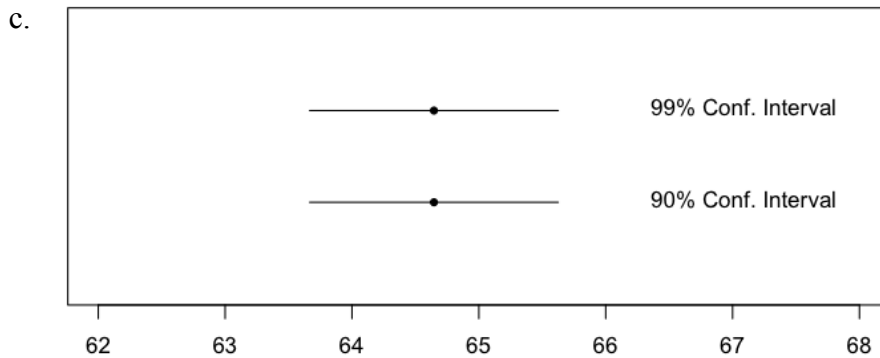
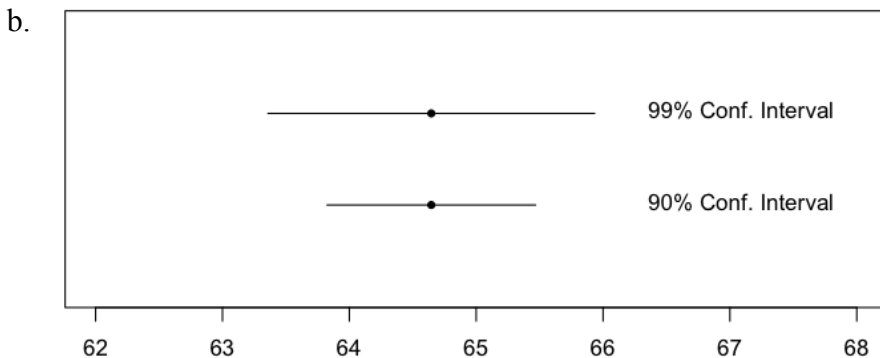
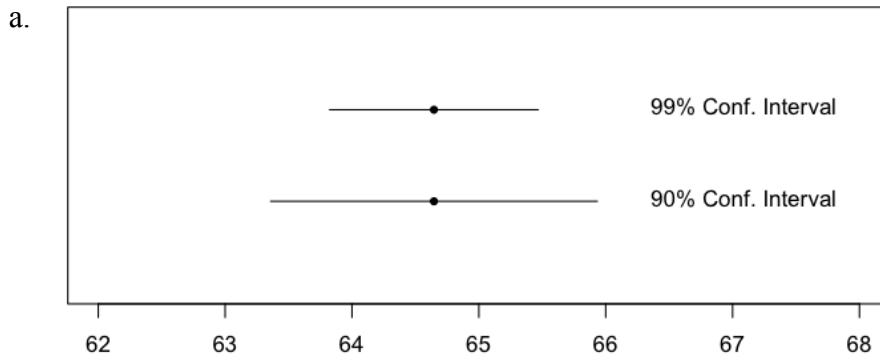
- a. The mean number of books that adults read during the year of 2011 was 16.564.
- b. The variability in the mean number of books from sample to sample is quite small spanning from approximately 15 to 18.
- c. The variability in the number of books from person to person is quite small spanning from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

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?
- Values approximately 17.2 and above because it is unlikely that adults would read that many books.
 - Values below approximately 15.0 and values above approximately 18.0 because there are no dots that are that extreme.
 - Values in the bottom 5% (below approximately 16.0) and values in the top 5% (above approximately 17.0).
 - Values in the bottom 2.5% (below approximately 15.7) and values in the top 2.5% (above approximately 17.4).
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?
- The average number of American adult cell phone users who access the internet on their phones in 2013.
 - The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
 - The percent of all American adult cell phone users who access the internet on their phones in 2013.
 - For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.
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%)?
- We know that 37% of veterans in the *sample* have been divorced at least once.
 - We know that 37% of veterans in the *population* have been divorced at least once.
 - We can say with 95% confidence that 37% of veterans in the *sample* have been divorced at least once.
 - We can say with 95% confidence that 37% of veterans in the *population* have been divorced at least once.

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?



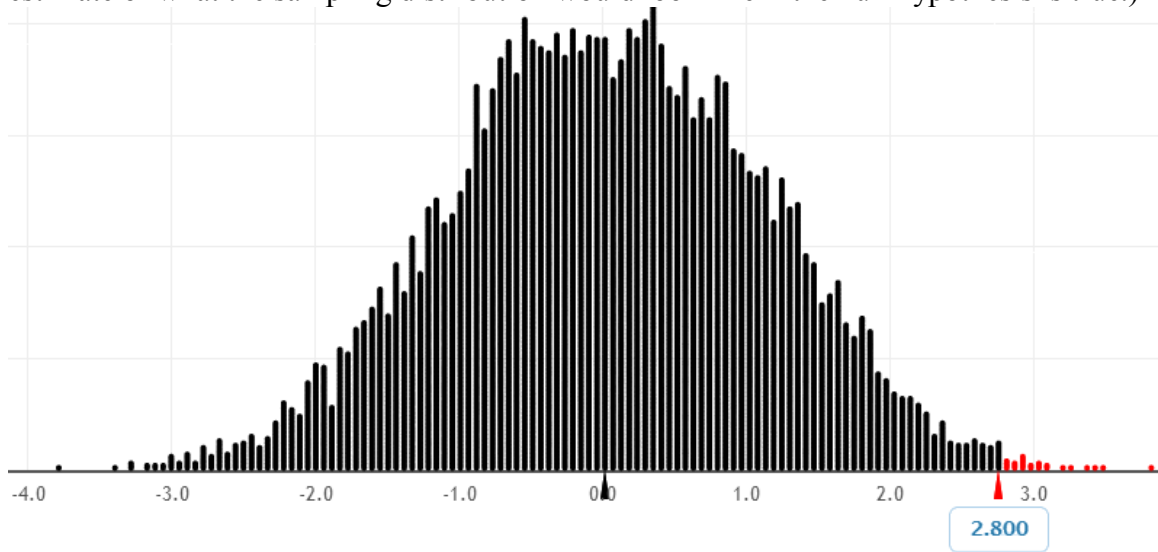
Items 23 and 24 refer to the following situation:

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)

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?
- No, because the average of the re-randomized sample mean differences is equal to 0.
 - No, because the proportion of re-randomized sample mean differences equal to or above 2.8 is very small.
 - Yes, because the proportion of re-randomized sample mean differences equal to or above 2.8 is very small.
 - Yes, because the observed result shows that the nap group remembered an average of 2.8 words more than the caffeine group.
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?
- Decrease, because with a larger sample size, there would be less variability in the re-randomized sample mean differences.
 - Increase, because with a larger sample size, there is more opportunity for error.
 - Stay about the same, because people are still being assigned to groups randomly.

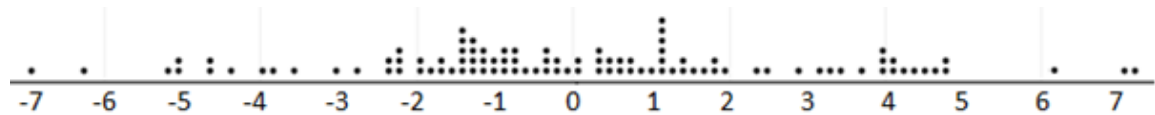
Items 25 and 26 refer to the following situation:

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 [$\text{mean}(\text{text message}) - \text{mean}(\text{control})$] 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)

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.

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.
- .03
 - .05
 - .06
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?
- The circuit is definitely not good and needs to be repaired.
 - The circuit is most likely not good, but it could be good.
 - The circuit is definitely good and does not need to be repaired.
 - The circuit is most likely good, but it might not be good.
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?
- The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%.
 - The researcher should conduct a hypothesis test because a hypothesis test is always appropriate.
 - The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur 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?”

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.
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.

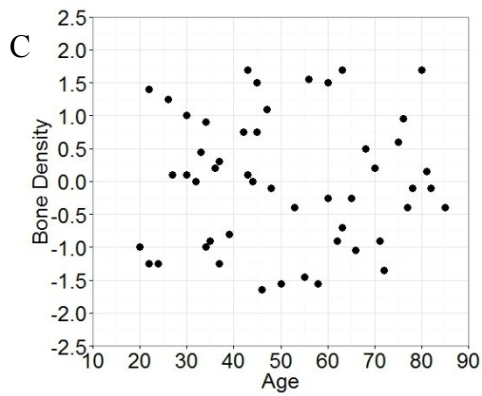
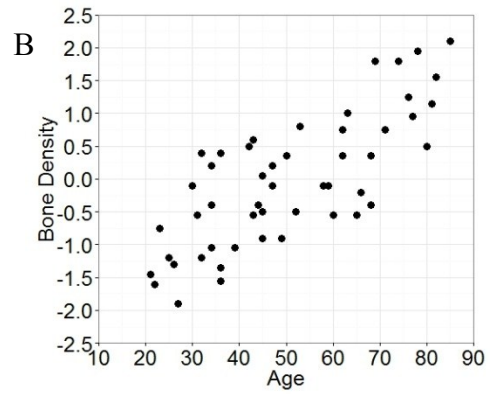
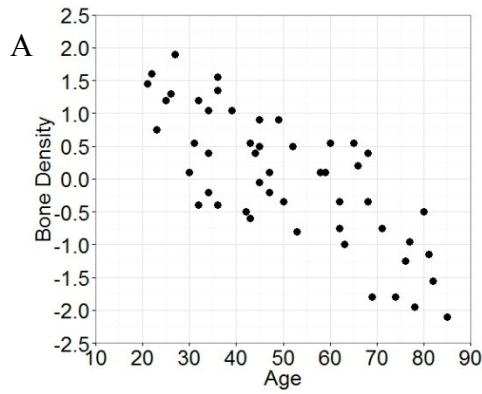
31. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of p -value would she want to obtain?
- A large p -value.
 - A small p -value.
 - The magnitude of a p -value has no impact on statistical significance.
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?
- Yes. It means you cannot conclude that the alternative hypothesis is true, so the null hypothesis must be true.
 - No. It means you cannot conclude that the null hypothesis is true, so the alternative hypothesis must be true.
 - No. It means that there is not enough evidence to conclude that the null hypothesis is false.
 - No. It means that there is not enough evidence to conclude that the alternative hypothesis is false.
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.
- Reject the null hypothesis and conclude that the dog correctly identifies cancer more than one fifth of the time.
 - There is enough statistical evidence to prove that the dog correctly identifies cancer more than one fifth of the time.
 - 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.

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?
- Observational study
 - Randomized experiment
 - Survey
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?

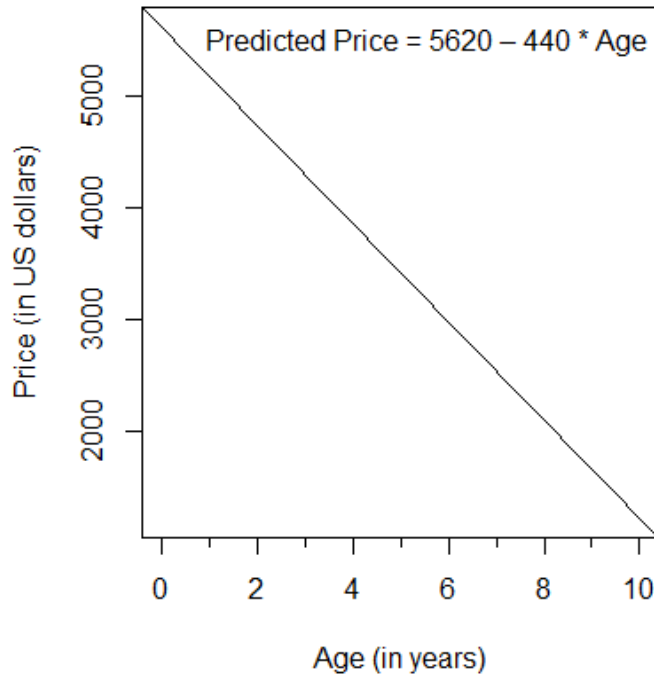
- Although 5,000 students live in dormitories on campus, only 500 were sent the survey.
- The survey was sent to only first-year students.
- Of the 500 students who were sent the survey, only 160 responded.
- All of the above present a problem for generalizing the results to all dormitory students at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

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



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

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?

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

Appendix D

Correspondence with Expert Reviewers of the Preliminary Assessment

D1 Invitation to Expert Reviewers of the Preliminary BLIS Assessment

Dear Professor XXX,

I am writing to request your assistance again in reviewing the first draft of the [*Basic Literacy In Statistics (BLIS) assessment*] for the introductory statistics course. As you know, 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. I am working with my co-advisers, Joan Garfield and Michelle Everson.

Thank you for reviewing the test blueprint and providing valuable feedback that led to its revision.

As stated in the test blueprint review, I am using this definition of statistical literacy: *being able to read, understand, and communicate statistical information.*

The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. This instrument includes items on modeling and simulation-based methods as well as some of the more traditional topics, because it is to be an assessment for a more modern version of the introductory course.

Your review will help provide evidence of the validity and value of the [BLIS assessment].

If you agree to do this, I would like you to take the perspective of an instructor who is teaching an introductory statistics course at the postsecondary level that includes modeling and simulation-based methods in the curriculum. You will be asked to rate how well the items align with the learning outcomes provided in the test blueprint, and your feedback will be used to modify the current [BLIS assessment] and create a final version of the [BLIS assessment]. Your feedback will be invaluable as I work toward creating a final version of this assessment.

I am attaching three documents:

- 1) a description of changes made to the test blueprint based on reviewer feedback,
- 2) a copy of the final test blueprint, and
- 3) the evaluation form that contains the preliminary version of the [BLIS assessment].

Ideally, I would like to receive a completed evaluation form from you within two weeks, by December 16, 2013. If this will not be possible, please let me know. Further, please

feel free to contact me with any questions as you evaluate the preliminary version of the [BLIS assessment]. I appreciate your time and expertise.

Please let me know as soon as possible if you are able to complete the review within two weeks.

Sincerely,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

D2 Summary of Changes to the Test Blueprint

Six reviewers, including you, provided their expertise in reviewing the preliminary test blueprint. Reviewers included leading statistics education researchers, introductory statistics instructors who teach using modeling and simulation-based methods, statistics assessment experts, and statistical literacy experts. This document provides a summary of the major changes made to the preliminary test blueprint.

Recall that the learning outcomes were rated on a four-point scale where a 1 represented a learning outcome “not essential for test” and a 4 represented a learning outcome “essential for test”. Most of the learning outcomes that had ratings with at least three 4’s or a combination of at least four 3’s and 4’s were kept in the test blueprint. Learning outcomes that had ratings with at least three 1’s were removed from the test blueprint.

Two methods were used to determine whether or not the remaining learning outcomes should be kept. First, comments from the reviewers were examined. In addition, the emphasis of the remaining learning outcomes in relatively new introductory statistics textbooks was used to help determine what learning outcomes should be kept and those that should not be kept.

Some reviewers commented that there was overlap in some of the learning outcomes so some learning outcomes were deleted to remove redundancy. For example, the following learning outcomes were similar: *Understanding that a confidence interval provides plausible values of the population parameter* and *Understanding of the purpose of a confidence interval*. Therefore, I deleted the second learning outcome.

Multiple reviewers said that bootstrap confidence intervals are not included in all introductory statistics courses that include modeling and simulation-based methods. As a result, I changed all learning outcomes that included bootstrap confidence intervals to confidence intervals in general. Items were written so that the confidence intervals were not specific to one method.

Reviewers were asked to suggest additional topics and learning outcomes. The definition of statistical literacy I am working with and the emphasis in relatively new introductory statistics textbooks were used to help determine what learning outcomes should be added to the test blueprint. Learning outcomes added to the test blueprint include: 1) *Ability to match a scatterplot to a verbal description of a bivariate relationship*, 2) *Ability to use a least squares regression equation to make a prediction*, and 3) *Understanding that every model is based on assumptions which limit our scope of inferences*.

D3 Preliminary Assessment Review Form

[Basic Literacy in Statistics (BLIS)] Review Form

The following table contains the items on the [BLIS assessment], which was designed for students enrolled in an introductory statistics course at the postsecondary level. Some of the items involve modeling and simulation-based methods because more and more of these topics are being included in the introductory statistics curriculum.

The [BLIS assessment] has 37 items compiled from different sources. Approximately half of the items were taken from existing exams including the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS), Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Topic Scale tests, and Goals and Outcomes Associated with Learning Statistics (GOALS). The remaining items are new items. All of the new items, except for item 3, are open-ended. After think-aloud interviews are conducted with students and a pilot test is run, the open-ended items will be converted to forced-choice items. I am open to changing any item, existing or new, based on reviewer feedback.

Your feedback on the [BLIS assessment] will help provide evidence that the assessment items measure the desired learning outcomes, which helps make an argument for the validity of the instrument.

In your review, please take the perspective of an instructor who is teaching an introductory statistics course at the postsecondary level that includes modeling and simulation-based methods (to some extent) in the curriculum.

Rating Procedure

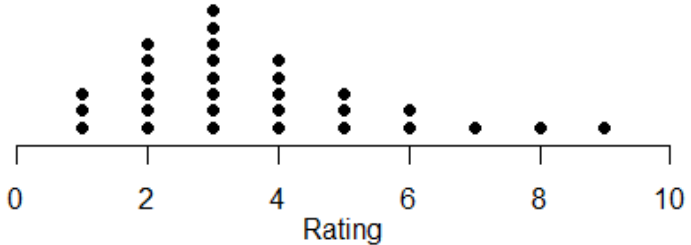
For each assessment item, rate the extent to which you agree or disagree that the item measures the intended learning outcome, as shown below:

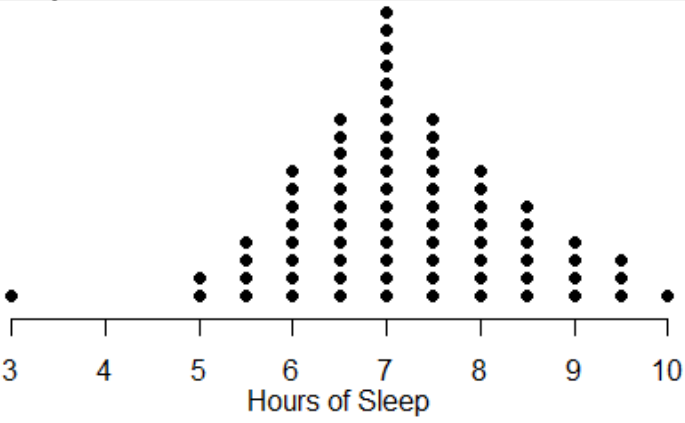
How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.	Strongly Disagree 1	Disagree 2	Agree 3	Strongly Agree 4
Learning Outcome: Understanding of the difference between a sample and population Assessment and Item #: New Item – Real-world context based on pewinternet.org/ 1. The Pew Research Center surveyed a nationally representative sample of 1,002 adults in 2013. The sample percent of internet users that have had an email or social networking account compromised was 21%. Identify the sample and population you would like to make inferences about.				
Comments (optional):				
Learning Outcome: Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term Assessment and Item #: New Item 2. Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 10 times and the other student flips a coin 50 times. Which student is more likely to get close to half of their coin flips heads up? Explain why you chose the student you did.				
Comments (optional):				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding that statistics computed from random samples tend to be centered at the parameter Assessment and Item #: New Item</p> <p>3. A manufacturer of frozen pizzas produces hamburger pizzas where the true average weight is 400 grams. To check the quality of the manufacturing process, random samples of 25 pizzas are taken daily and the average weight is recorded. Assuming that nothing is wrong with the manufacturing process, which sequence below is the most plausible for the average weight for five samples?</p> <p>a. 381, 389, 405, 424, 441. b. 336, 362, 377, 387, 400. c. 395, 402, 420, 445, 450. d. Any of the above.</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Ability to determine what type of study was conducted Assessment and Item #: ARTIST Topic Scale Test – Data Collection 7 – Modified</p> <p>4. Suppose a researcher wanted to determine if aspirin reduces the chance of a heart attack. The researcher studied 500 patients that 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?</p> <p>a. Observational b. Experimental c. Survey d. None of the above</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>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 = subcompact, 2 = compact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.</p> <p>Learning Outcome: Ability to determine if a variable is quantitative or categorical Assessment and Item #: ARTIST Topic Scale Test – Data Collection 2</p> <p>5. What type of variable is this?</p> <ul style="list-style-type: none"> a. categorical b. quantitative c. continuous 				
<p>Comments (optional):</p>				
<p>Learning Outcome: Ability to determine if a variable is an explanatory variable or a response variable Assessment and Item #: ARTIST Topic Scale Test – Data Collection 3</p> <p>6. The student plans to see if there is a relationship between the number of speeding tickets a student gets in a year and the type of vehicle he or she drives. Identify the response variable in this study.</p> <ul style="list-style-type: none"> a. college students b. type of car c. number of speeding tickets d. average number of speeding tickets last year 				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding of the difference between a statistic and parameter Assessment and Item #: ARTIST Website Item ID = Q0618</p> <p>7. CNN conducted a quick vote poll on September 19, 1999 to determine “What proportion of Americans still think that the Miss America pageant is still relevant today?” The poll was conducted on the internet. Here are the results of the poll: Is the Miss American pageant still relevant today? Yes: 1192 votes, No: 4389 votes; Total: 5581 votes. Describe the parameter of interest.</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understanding that statistics vary from sample to sample Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)</p> <p>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. Researcher B claims that the first study must have been flawed because the mean was not the same in both studies. How would you respond to Researcher B’s statement?</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Ability to describe and interpret a dotplot Assessment and Item #: ARTIST Topic Scale Test – Data Representation 12 – Modified</p> <p>9. One of the items on the student survey for an introductory statistics course was "Rate your aptitude to succeed in this class on a scale of 1 to 10" where 1 = Lowest Aptitude and 10 = Highest Aptitude. The instructor examined the data for men and women separately. Below is the distribution of this variable for the 30 women in the class.</p>  <p>How should the instructor interpret the women's perceptions regarding their success in the class?</p> <ol style="list-style-type: none"> A majority of women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding. The women in the class see themselves as having lower aptitude for statistics than the men in the class. If you remove the three women with the highest ratings, then the result will show an approximately normal distribution. 				
<p>Comments (optional):</p>				

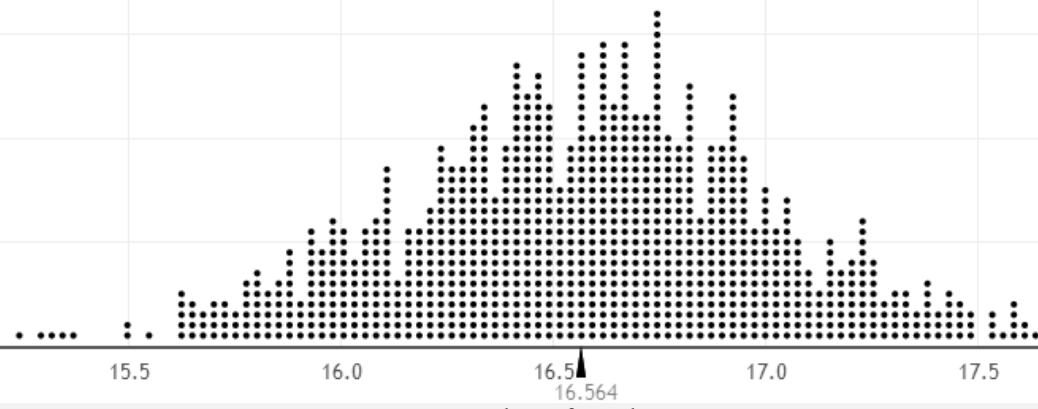
<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data Assessment and Item #: CAOS 1 – Modified</p> <p>10. The following graph shows a distribution of hours slept last night by a group of college students.</p>  <p>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.</p> <ol style="list-style-type: none"> The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five. The distribution is normal, with a mean of about 7 and a standard deviation of about 1. Most students seem to be getting enough 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. The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours. 				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding the importance of creating graphs prior to analyzing data Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)</p> <p>11. A researcher was interested in determining if a smaller dose of a particular flu vaccine was just as successful as a full dose. A randomized experiment was conducted with 100 participants. Half of the participants received the full dose of the vaccine and the other half received a half does of the vaccine. The number of days the participant got the flu 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 get the flu for the full dose group and half dose group. What step should the research take after the data is collected but before the hypothesis test is conducted?</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Ability to interpret a percent in the context of the data Assessment and Item #: ARTIST Topic Scale Test – Probability 9</p> <p>12. The local Meteorological claims that there is a 70% probability of rain tomorrow. Provide the best interpretation of this statement.</p> <p>a. Approximately 70% of the city will receive rain within the next 24 hours. b. Historical records show that it has rained on 70% of previous occasions with the same weather conditions. c. If we were to repeatedly monitor the weather tomorrow, 70% of the time it will be raining. d. Over the next ten days, it should rain on seven of them.</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Ability to interpret a mean in the context of the data Assessment and Item #: New Item</p> <p>13. According to a pet store owner, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean.</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understand how a mean is affected by skewness or outliers Assessment and Item #: ARTIST Topic Scale Test – Sampling Variability 7</p> <p>14. The distribution for a population of measurements is presented below.</p> <div data-bbox="340 675 833 984" data-label="Figure"> </div> <p>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?</p> <ul style="list-style-type: none"> a. 4 to 6 b. 7 to 9 c. 10 to 12 				
<p>Comments (optional):</p>				

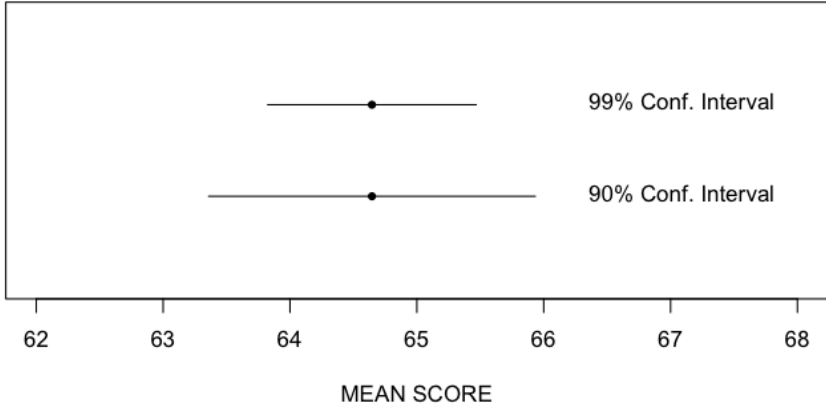
<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Ability to interpret a standard deviation in the context of the data Assessment and Item #: ARTIST Topic Scale Test – Measures of Spread 2</p> <p>15. The 30 introductory statistics students took another quiz worth 30 points. On this quiz, the standard deviation of the scores of that quiz was 1 point. Which of the following gives the most suitable interpretation?</p> <ul style="list-style-type: none"> a. all of the individual scores are one point apart b. the difference between the highest and lowest score is 1 c. the difference between the upper and lower quartile is 1 d. a typical score is within 1 point of the mean 				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understanding of the properties of standard deviation Assessment and Item #: ARTIST Topic Scale Test – Measures of Spread 5</p> <p>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 for the class to be -2.30. What do we know?</p> <ul style="list-style-type: none"> 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. 				
<p>Comments (optional):</p>				

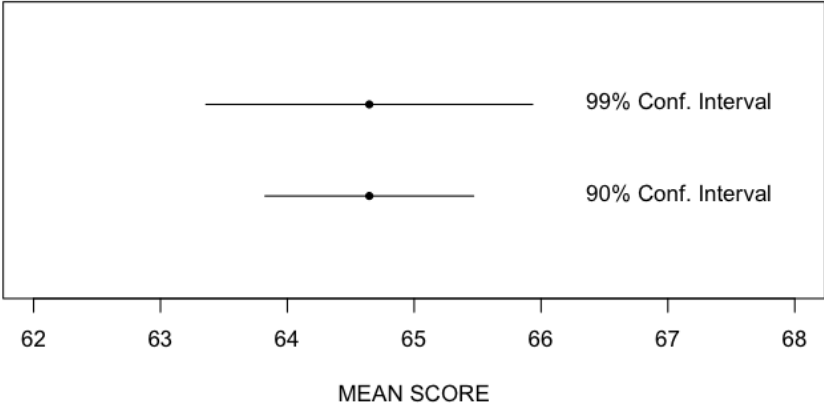
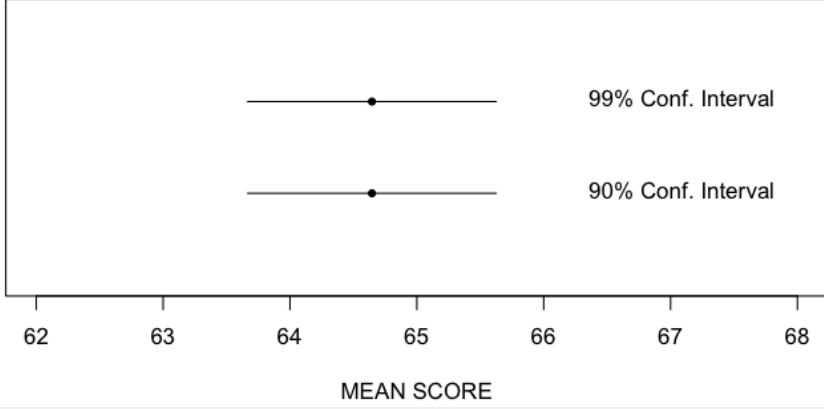
<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding of what an empirical sampling distribution represents Assessment and Item #: ARTIST Topic Scale Test – Sampling Variability 1 – Modified</p> <p>17. Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights of a random sample of 3 pebbles each, with the mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.</p> <div style="display: flex; justify-content: space-around;"> <div data-bbox="289 667 787 919"> <p>A</p> </div> <div data-bbox="814 667 1312 919"> <p>B</p> </div> </div> <p>a. No, in both Figure A and Figure B, the X represents one pebble that weights 6 grams. b. Yes, Figure A has a larger range of values than Figure B. c. Yes, the X in Figure A is the weight for a single pebble, while the X in Figure B represents the average weight of 3 pebbles.</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>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 graphed by doing the following:</p> <ul style="list-style-type: none"> • 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. <p>Below is the plot of the empirical sampling distribution. (Note: This plot can be used as an estimate of what the sampling distribution would look like.)</p> 				

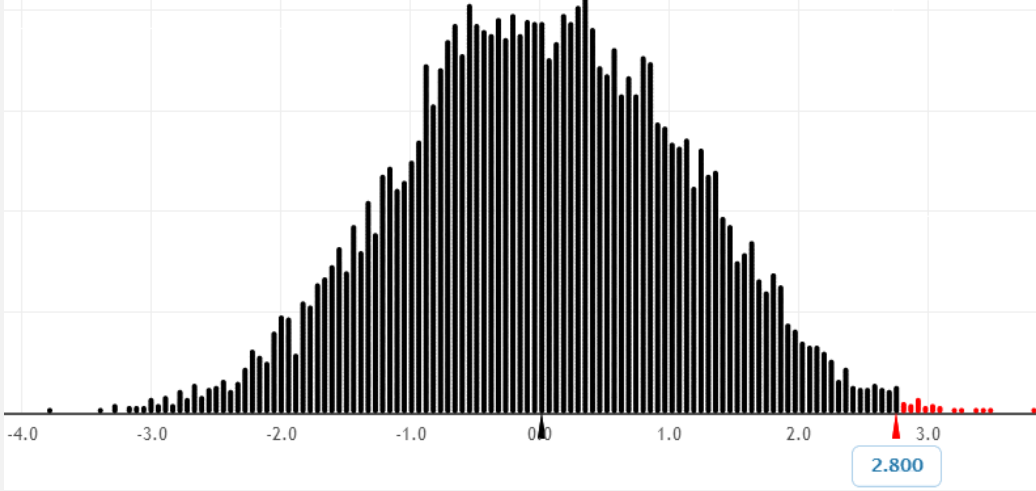
<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding that an empirical sampling distribution shows how sample statistics tend to vary Assessment and Item #: New Item – Real-world context based on pewinternet.org/</p> <p>18. What information is obtained from this distribution?</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter Assessment and Item #: New Item – Real-world context based on pewinternet.org/</p> <p>19. What values do you believe would NOT be plausible estimates of the population average number of books read? Explain your answer.</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understanding that a bootstrap interval provides plausible values of the population parameter Assessment and Item #: New Item – Real-world context based on pewinternet.org/</p> <p>20. The Pew Research Center surveyed 2,076 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% bootstrap interval was 58% to 62%. What is this interval attempting to estimate?</p>				
<p>Comments (optional):</p>				

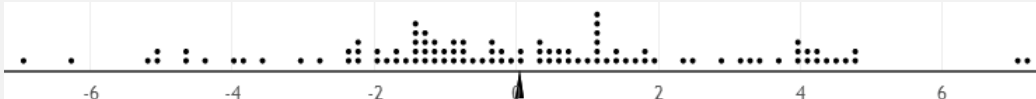
<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding that a confidence interval is centered at the sample statistic Assessment and Item #: ARTIST TestBank Item ID = Q0943 – Modified</p> <p>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%)?</p> <ul style="list-style-type: none"> 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 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 				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding of how the confidence level affects the width of a confidence interval Assessment and Item #: GOALS 14 – Modified</p> <p>22. This question 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.</p> <p>Consider a standardized test that has been given to thousands of high school students.</p> <p>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 <i>and</i> a 90% confidence interval for the population mean are constructed using this new sample.</p> <p>Which of the following options would best represent how the two confidence intervals would compare to each other?</p> <p>a.</p>  <p>The graph displays two horizontal lines representing confidence intervals on a horizontal axis labeled 'MEAN SCORE' with tick marks at 62, 63, 64, 65, 66, 67, and 68. The upper line is labeled '99% Conf. Interval' and has a solid dot at 65. The lower line is labeled '90% Conf. Interval' and also has a solid dot at 65. The 99% interval is narrower than the 90% interval.</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>b.</p>  <p>c.</p> 				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>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.</p> <p>A randomization distribution was graphed by doing the following:</p> <ul style="list-style-type: none"> • From the original sample, the 24 participants were re-randomized to the nap group (n=12) and caffeine group (n=12), without replacement. • The mean difference in words recalled was computed for the re-randomized groups and placed on the plot shown below. • This was repeated 999 more times. <p>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.)</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
 <p>(Words Recalled for Nap Group) – (Words Recalled for Caffeine Group)</p> <p>Learning Outcome: Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)</p> <p>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 for the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis? Explain your answer.</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding of how sample size affects the standard error Assessment and Item #: New Item – Real-world context based on example from Gould and Ryan (2013)</p> <p>24. Suppose the sample size was doubled from 24 participants to 48 participants and the participants were still split equally into the two groups. Would the standard error change? If so, how? Explain your answer.</p>				
<p>Comments (optional):</p>				
<p>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. 26 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. The average weight loss for the text message group was 2.8 pounds and -2.6 pounds for the control group. Note that the control group had a negative weight loss which means that they actually gained weight. The difference in average weight loss was $(2.8) - (-2.6) = 5.4$.</p> <p>A randomization distribution was graphed by doing the following:</p> <ul style="list-style-type: none"> From the original sample, the 50 women were re-randomized to the text message group (n=26) and control group (n=24), without replacement. The mean difference in weight loss was computed for the re-randomized groups and placed on the plot shown below. This was repeated 99 more times. <p>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.)</p>  <p>(Weight Loss for Text Message Group) – (Weight Loss for Control Group)</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding that a randomization distribution tends to be centered at the hypothesized null value Assessment and Item #: New Item – Real-world context based on Steinberg, Levine, Askew, Foley, and Bennett (2013)</p> <p>25. Why is the randomization distribution centered at 0?</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Ability to estimate a p-value using a randomization distribution Assessment and Item #: New Item – Real-world context based on Steinberg, Levine, Askew, Foley, and Bennett (2013)</p> <p>26. The alternative hypothesis is that the text messages lead to a higher weight loss than no text messages for women participating in this weight loss program. Therefore a one-tailed (i.e., one-sided) test will be conducted. Compute the p-value for the observed difference in mean weight loss of 5.4 based on the simulated data. Show how to find this p value and explain each step.</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding of the logic of a hypothesis test Assessment and Item #: CAOS 40</p> <p>27. The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?</p> <p>a. The circuit is definitely not good and needs to be repaired. b. The electrician decides that the circuit is defective, but it could be good. c. The circuit is definitely good and does not need to be repaired. d. The circuit is most likely good, but it could be defective.</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understanding of the purpose of a hypothesis test Assessment and Item #: New Item – Real-world context based on Chervin et al. (2013)</p> <p>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)? 20 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. 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 statistics (70%) as evidence of the effectiveness? Explain your answer.</p>				
<p>Comments (optional):</p>				

How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.	Strongly Disagree 1	Disagree 2	Agree 3	Strongly Agree 4
<p>Learning Outcome: Understanding that every model is based on assumptions which limit our scope of inferences Assessment and Item #: New Item</p> <p>29. A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions as a model to make inferences about the two states? Explain your answer.</p>				
Comments (optional):				
<p>Learning Outcome: Ability to determine a null and alternative hypothesis statement based on a research question Assessment and Item #: New Item – Real-world context based on www.wilderresearch.org/</p> <p>30. 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 days and 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 is difference for males and females with regards to the average number of nights spent in a place not intended for housing?” In order to conduct a hypothesis test to answer this research question, what would the null and alternative hypothesis statements be?</p>				
Comments (optional):				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Ability to determine statistical significance based on a p-value Assessment and Item #: CAOS 19</p> <p>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 <i>p</i>-value would she want to obtain?</p> <p>a. A large <i>p</i>-value. b. A small <i>p</i>-value. c. The magnitude of a <i>p</i>-value has no impact on statistical significance.</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understanding that errors can occur in hypothesis testing Assessment and Item #: New Item – Real-world context based on example from Utts (2003)</p> <p>32. A clinical trial was conducted to determine if women who have regular mammograms to screen for breast cancer would describe 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 as 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? Explain your answer.</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding of how a significance level is used Assessment and Item #: New Item – Real-world context based on example from Lock, Lock, Lock, Lock, and Lock (2013)</p> <p>33. 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 research question is “Does the dog correctly identify cancer more than half of the time?” The p-value is less than .001. Using a significance level of .05, what conclusion should be made? Explain why you chose to make your conclusion.</p>				
<p>Comments (optional):</p>				
<p>Learning Outcome: Understanding that only an experimental design with random assignment can support causal inference Assessment and Item #: ARTIST Topic Scale Test – Data Collection 4</p> <p>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?</p> <ul style="list-style-type: none"> a. Correlational study b. Randomized experiment c. Time Series study d. Survey 				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Understanding of the factors that allow a sample of data to be generalized to the population Assessment and Item #: CAOS 38</p> <p>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. Which of the following does NOT affect the college official's ability to generalize the survey results to all dormitory students?</p> <ul style="list-style-type: none"> a. Five thousand students live in dormitories on campus. A random sample of 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. 				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Ability to match a scatterplot to a verbal description of a bivariate relationship Assessment and Item #: GOALS 7</p> <p>36. Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?</p> <div data-bbox="363 527 1228 1198"> <p>Graph A: Scatterplot showing a negative correlation between AGE and Bone Density. The data points generally trend downwards from left to right.</p> <p>Graph B: Scatterplot showing a positive correlation between AGE and Bone Density. The data points generally trend upwards from left to right.</p> <p>Graph C: Scatterplot showing a non-linear relationship between AGE and Bone Density. The data points form a U-shape, with lower bone density values at intermediate ages and higher values at the extremes.</p> </div> <p>a. Graph A b. Graph B c. Graph C</p>				
<p>Comments (optional):</p>				

<p>How much you agree or disagree with the following statement? The assessment item measures the specified learning outcome.</p>	<p>Strongly Disagree 1</p>	<p>Disagree 2</p>	<p>Agree 3</p>	<p>Strongly Agree 4</p>
<p>Learning Outcome: Ability to use a least-squares regression equation to make a prediction Assessment and Item #: ARTIST Topic Scale Test – Bivariate Data, Quantitative 13</p> <p>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 model:</p> <p>Predicted Price = $5620 - 440 * \text{Age}$</p> <p>A friend asked him to predict the price of a 5 year old model of this car, using his equation. Which of the following is the most correct response to provide?</p> <ol style="list-style-type: none"> Plot a regression line, find 5 on the horizontal axis, and read off the corresponding value on the y axis. Substitute 5 in the equation and solve for "price". Both of these methods are correct. Neither of these methods is correct. 				
<p>Comments (optional):</p>				

- Please add any suggestions you have for improving the [BLIS assessment].

Thank you for helping develop the [BLIS assessment].

Appendix E

Correspondence with Instructors and Students for the Cognitive Interviews

E1 In-class Invitation Script

Hello students.

You are invited to participate in a research study designed to develop and provide validity evidence for a research instrument called the [*Basic Literacy In Statistics (BLIS) assessment*]. The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level.

This study is being conducted by me, a doctoral candidate in the Department of Educational Psychology under the supervision of Dr. Joan Garfield.

You will participate in a one-hour interview that is designed to gain an understanding of what reasoning and strategies you use for the questions in the [BLIS] assessment. You will be asked to talk aloud as you solve the problems. You will also be asked to say whatever you are looking at, thinking, doing, and feeling as you take the assessment. You will be audio-taped to produce a record of your responses for later analysis.

The problems may not look like anything you have done before and a problem may have a solution that you can produce using everyday knowledge and reasoning. While the test will cover some of what you learned in your statistics course, you do not have to review the course content for this study.

As an incentive to participate in this study, you will receive a \$20 *Amazon.com* gift certificate.

I am planning to conduct the interviews from December 30th to January 12th. 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. After a time is set up, you will be told the time and location of the study.

You will be notified by December 20th 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.

E3 Email Invitation to Students for the In-person Cognitive Interviews

To: Students who have taken in introductory statistics course

You are invited to participate in a research study designed to develop and provide validity evidence for a research instrument called the [*Basic Literacy In Statistics (BLIS) assessment*]. As an incentive to participate in this study, you will receive a \$20 *Amazon.com* gift certificate.

The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. Statistical literacy skills include being able to read, understand, and communicate statistical information. You were selected as a possible participant because you are currently taking or have taken a postsecondary introductory statistics course.

This study is being conducted by Laura Ziegler, a doctoral candidate in the Department of Educational Psychology under the supervision of Dr. Joan Garfield.

If you agree to participate, you will be asked take part in a one-hour interview that is designed to gain an understanding of the kinds of reasoning and strategies you use to answer the questions in the [BLIS] assessment. You will be asked to talk aloud as you solve the problems. You will also be asked to say whatever you are looking at, thinking, doing, and feeling as you take the assessment. You will be audio-taped to produce a record of your responses for later analysis.

Interviews will take place late December or early January.

If you are interested in participating, please email Laura Ziegler at sath0166@umn.edu by December 29th. Please indicate the days and times you will be free (for approximately one hour) from December 30th to January 12th. If you want to participate and you are not available during these days, please contact Laura to find out if another day might work.

You will be notified by December 30th if you are selected to participate in the study.

Thank you in advance for your time and your consideration of this request!

E4 Email Invitation to Students for the Skype Cognitive Interviews

To: Students who have taken in introductory statistics course

You are invited to participate in a research study designed to develop and provide validity evidence for a research instrument called the [*Basic Literacy In Statistics (BLIS) assessment*]. As an incentive to participate in this study, you will receive a \$20 *Amazon.com* gift certificate.

The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. Statistical literacy skills include being able to read, understand, and communicate statistical information. You were selected as a possible participant because you have taken a postsecondary introductory statistics course.

This study is being conducted by Laura Ziegler, a doctoral candidate in the Department of Educational Psychology under the supervision of Dr. Joan Garfield.

If you agree to participate, you will be asked take part in a one-hour interview via Skype that is designed to gain an understanding of the kinds of reasoning and strategies you use to answer the questions in the [BLIS] assessment. You will be asked to talk aloud as you solve the problems. You will also be asked to say whatever you are looking at, thinking, doing, and feeling as you take the assessment. You will be audio-taped to produce a record of your responses for later analysis.

Interviews will take place early February.

If you are interested in participating, please email Laura Ziegler at sath0166@umn.edu. Please indicate the days and times you will be free (for approximately one hour) from now to January 21st including weekends.

You will be notified if you are selected to participate in the study.

Thank you in advance for your time and your consideration of this request!

Sincerely,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

E5 Letter Sent to Instructors Requesting Them to Email the Interview Invitation to Their Students

Dear Fellow Stat Chatters,

Do you teach an introductory statistics course that includes some simulations? If yes, I am writing to see if you might forward a message from me to your students. I am developing a new instrument for my dissertation research: the [*Basic Literacy In Statistics (BLIS) assessment*]. I am working with my co-advisers, Joan Garfield and Michelle Everson.

I am looking for two students who would be willing to be interviewed as I ask them to read and respond to the questions on the assessment. As an incentive to students, they will receive a \$20 Amazon.com gift certificate.

The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. I define statistical literacy as being able to read, understand, and communicate statistical information. This instrument includes items on modeling and simulation-based methods as well as some of the more traditional topics, because it is to be an assessment for a more modern version of the introductory course.

Your students' feedback will be invaluable as I work toward creating a final version of this assessment. Please let me know if you are willing to send the email below by replying to this email (sath0166@umn.edu). In addition, please let me know which textbook you teach your course with.

Sincerely,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

E6 Cognitive 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 and be permitted to leave.

E7 Consent Form for Cognitive Interviews

This study is being conducted by a researcher from the University of Minnesota. You are invited to participate in a research study designed to develop and provide validity evidence for a research instrument called the [*Basic Literacy In Statistics (BLIS) assessment*]. You were selected as a possible participant because you are currently taking or have taken a postsecondary introductory statistics course. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by Laura Ziegler, a doctoral candidate in the Department of Educational Psychology under the supervision of Dr. Joan Garfield.

Background Information

The proposed study is to develop an instrument to assess college students' statistical literacy skills. The target population of the instrument is college students in the U.S. who are taking a non-calculus-based introductory statistics course that includes modeling and simulation-based methods. The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. Statistical literacy skill include being able to read, understand, and communicate statistical information.

Procedures

You will participate in a one-hour interview that is designed to gain an understanding of what reasoning and strategies you use for the questions in the [BLIS] 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 skills. These excerpts may be in the form of a transcription of your statements during the interview, or of audio files selected from an interview.

We are asking for your consent to do three things. First, we ask for your consent to audio-tape and record the interview. Second, we ask for your consent to include audio files of your interviews in presentations of this research. Third, we ask for your consent to include excerpts of your statements during the interviews in research presentations and publications.

Compensation

You will receive a \$20 *Amazon.com* gift certificate for your participation in the one-hour interview.

Risks and Benefits of Being in the Study

There are no known risks to you as a participant.

The benefit to participate is the opportunity to develop a better understanding of statistics, and of your own statistical literacy knowledge.

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 de-identified and stored on a password protected computer; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study

Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota. 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 Laura Ziegler under the advisement of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Michelle Everson Ph.D. (Educational Psychology – Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Laura Ziegler, at sath0166@umn.edu. You may also contact my advisor, Dr. Joan Garfield, at jbg@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.

You will be given a copy of this form to keep for your records.

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 audio files of my interview in presentations of this research.

I give permission to include excerpts of my statements in research presentations and publications.

Your Name (Please PRINT): _____

Signature: _____ Date: _____

Appendix F

Correspondence with Instructors and Students for the Pilot Test

F1 Invitation to Instructors to Participate in the Pilot Test

Dear Professor XXX,

I am writing to request your assistance in helping me pilot the [*Basic Literacy In Statistics (BLIS) assessment*] for the introductory statistics course. 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. I am working with my co-advisers, Joan Garfield and Michelle Everson.

I am requesting your help with this project because you teach introductory statistics course at the postsecondary level that includes modeling and simulation-based methods in the curriculum.

I define statistical literacy as being able to read, understand, and communicate statistical information. The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. This instrument includes items on modeling and simulation-based methods as well as some of the more traditional topics, because it is to be an assessment for a more modern version of the introductory course.

In light of changes to the introductory statistics course, namely the move to include modeling and simulation-based methods, a new assessment of statistical literacy is needed. A new statistical literacy assessment could be used for a variety of purposes. For example, the assessment could be used as a pretest to determine what statistical literacy skills students have prior to taking an introductory statistics course at the postsecondary level. Statistics instructors could also use the assessment in their classes to see which statistical literacy topics students understand and which topics they do not understand. Further, researchers could use the assessment to determine how much statistical literacy students gain in an introductory statistics course that incorporates modeling and simulation-based methods in the curriculum, or to determine if the amount of statistical literacy gained in introductory statistics courses taught with varying teaching methods or curricula is different.

In order to provide evidence of the validity and reliability of the [BLIS assessment], I am requesting your help at this time to pilot the preliminary instrument.

If you agree to do this, you will learn more about your students' statistical literacy knowledge. You can choose to administer the [BLIS assessment] to your student either in or outside of the classroom. You can choose to provide extra credit to your students or give no credit. You will receive your students' scores after they complete the exam.

I am attaching the preliminary version of the [BLIS assessment] for you to consider.

Ideally, I would like you to administer the [BLIS assessment] to your students by March 7, 2014. 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.

I hope you will agree to administer the preliminary version of the [BLIS assessment] to your students. Please let me know your response by replying to this email (sath0166@umn.edu).

Sincerely,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

F2 Instructions for Instructors on How to Administer the BLIS-2 Assessment to Their Students

Dear Professor XXX,

Thank you again for agreeing to administer the [Basic Literacy In Statistics assessment] to your students. Students can take the assessment online at:
https://umn.qualtrics.com/SE/?SID=SV_08mJJDKuGVplzut
Below is detailed information for administering the assessment in your statistics course.

First, we do not expect your students to complete the [Basic Literacy In Statistics assessment] in a proctored classroom. However, please instruct your students that they are expected to not work together on the test.

Second, while assessment is not a timed test, please allow at least 60 minutes for students to take the test. If students take the test out-of-class, let them know that we expect most students to take about 50 to 70-minutes to complete the test.

Third, at the beginning of the online test, students must provide their first and last names. They are then asked to enter a unique code that identifies your class. The code for your class is: CODE
Please provide this code (CODE) to your students along with the online link for the test. Students will then sign an online consent form.

After students complete the test, they will be asked to provide some additional information that is optional (e.g., gender and ethnicity).

Finally, in order to gather information on your course characteristics, I would like to ask you to complete a 30-second survey before or after you administer the [BLIS] test (the survey link:
https://umn.qualtrics.com/SE/?SID=SV_0SuWHtz7AWSXTDv). All identifying information from this survey will be removed from your record and only aggregate data will be presented in the sharing of results through publications or talks. Your participation is very much appreciated and is completely voluntary.

Please send me an email when your students have completed the assessment so I can send you an Excel file of the student scores.

Please be assured that your participation is very important to provide validity and reliability evidence for the proposed assessment. I really appreciate your help in attaining a broad sample of students.

In summary:

Link to the test: https://umn.qualtrics.com/SE/?SID=SV_08mJJDKuGVplzut

Unique class code: CODE

Link to the instructor survey:

https://umn.qualtrics.com/SE/?SID=SV_0SuWHtz7AWSXTDv

Thank you,

Laura Ziegler

Doctoral Candidate

Department of Educational Psychology

University of Minnesota

F3 Instructor Survey for the Pilot Test

1. What is your name?
2. What is the name of your institution?
3. What state is your institution in? (dropdown list included in online version of survey)
4. What type of institution are you administering the assessment at?
 - 2-year/technical college
 - 4-year college
 - University
5. What mathematics prerequisites are there for the course that you are administering the assessment in? (selected all that apply)
 - No mathematics requirement
 - High school algebra
 - College algebra
 - Calculus
6. Will the assessment be administered in class or out of class?
 - In class
 - Out of class
7. Will the assessment scores be used as a homework, quiz, or exam score for students?
 - Yes
 - No
8. Will the assessment scores be used to give extra credit to students?
 - Yes
 - No

F4 Consent Form for the Pilot Test and Instructions for Students on How to Complete the BLIS-2 Assessment

*Please read the description below and check in the Statement of Consent if you agree to participate in this study. * This question is required.*

This study is being conducted by a researcher from the University of Minnesota. You are invited to participate in a research study designed to develop and provide validity evidence for a research instrument called the [*Basic Literacy In Statistics (BLIS) assessment*]. You were selected as a possible participant because you are currently taking a postsecondary introductory statistics course. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by Laura Ziegler, a doctoral candidate in the Department of Educational Psychology under the supervision of Dr. Joan Garfield.

Background Information

The proposed study is to develop an instrument to assess college students' statistical literacy skills. The target population of the instrument is college students in the U.S. who are taking a non-calculus-based introductory statistics course that includes modeling and simulation-based methods. The purpose of the [BLIS assessment] is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. Statistical literacy skill include being able to read, understand, and communicate statistical information.

Procedures

If you agree to be in this study, you will complete an online version of the assessment. The assessment consists of 37 questions and will take 50 to 70 minutes to complete.

Risks and Benefits of Being in the Study

There are no known risks to you as a participant.

The benefit to participate is the opportunity to develop a better understanding of statistics, and of your own statistical literacy knowledge. 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 de-identified and stored on a password protected computer; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study

Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota. 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 Laura Ziegler under the advisement of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Michelle Everson Ph.D. (Educational Psychology – Statistics Education). If you have any questions you are encouraged to contact me, Laura Ziegler, at sath0166@umn.edu. You may also contact my advisor, Dr. Joan Garfield, at jbg@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 ASL online test. This test includes 37 multiple-choice types of questions. Please read each question carefully. 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.

F5 Demographic Questions for Students Taking the BLIS-2 Assessment

The following questions are optional.

1. What is your gender?
 - Female
 - Male
2. What age category do you fall in?
 - 19 or younger
 - 20 to 24
 - 25 to 29
 - 30 to 34
 - 35 to 39
 - 40 to 44
 - 45 to 49
 - 50 to 54
 - 55 or older
3. What is your class level?
 - Freshman/first year
 - Sophomore
 - Junior
 - Senior
 - Graduate student
 - Other
4. Are you an international student or foreign national student?
 - Yes
 - No
5. What is your racial or ethnic identification? (select all that apply)
 - White
 - Black or African American
 - American Indian or Alaska Native
 - Asian
 - Pacific Islander
 - Other

Appendix G

Correspondence with Instructors and Students for the Field Test

G1 Invitation Letter to Instructors to Participate in the Field Test Sent Via Personal Emails and the Isolated Statisticians Electronic Mailing List

Dear Instructor,

Do you teach an introductory statistics course at the postsecondary level that includes simulation, to some extent, in the curriculum?

If yes, I am writing to request your assistance in helping me to gather some pilot data to assist in the development a new instrument: *Basic Literacy in Statistics (BLIS) Assessment*. 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. I am working with my co-advisers, Joan Garfield and Michelle Everson.

For the purpose of this instrument, statistical literacy is defined as being able to read, understand, and communicate statistical information. The purpose of the BLIS Assessment is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. Because it is to be used as an assessment in a more modern version of the introductory course, this instrument includes items on simulation-based methods as well as some of the more traditional topics.

In light of changes to the introductory statistics course, namely the move to include simulation-based methods, a new assessment of statistical literacy is needed. A new statistical literacy assessment may be used for a variety of purposes. For example, the assessment could be used as a pretest to determine what statistical literacy skills students have prior to taking an introductory statistics course at the postsecondary level. Statistics instructors could also use the assessment in their classes to see which statistical literacy topics students understand and which topics they do not understand. Further, researchers can use the assessment to determine how much statistical literacy students gain in an introductory statistics course that incorporates simulation-based methods in the curriculum, or to determine if the amount of statistical literacy gained in introductory statistics courses taught with varying teaching methods or curricula is different.

I am requesting your help at this time to administer the instrument to your students. The data gathered will be used to evaluate the validity and reliability of the BLIS Assessment.

If you agree to do this, I will provide summary data to you to describe the statistical literacy of your students as well as how they compare to students at other institutions. You can choose to administer the BLIS Assessment to your students either inside or outside of the classroom. You may want to provide extra credit to your students as an

incentive. You will receive your students' scores after they complete the exam. Upon request, I will provide a summary of the results collected from other institutions.

Ideally, I would like you to administer the BLIS Assessment to your students by April 25, 2014. 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.

Please let me know if you are interested by emailing me at sath0166@umn.edu. I will then send you a copy of the BLIS Assessment and additional information for your consideration.

Sincerely,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

G2 Invitation Letter to Instructors to Participate in the Field Test Sent Via the Consortium for the Advancement of Undergraduate Statistics Education Mailing List

Dear Instructor,

I am writing to request your assistance in helping me to gather some pilot data to assist in the development a new assessment of statistical literacy. Below are the details. Thank you for your consideration.

Laura Ziegler, UMN

I am developing a new instrument, the *Basic Literacy in Statistics (BLIS) Assessment*, to be used in an introductory statistics course at the postsecondary level that includes simulation, to some extent, in the curriculum. This instrument is being developed 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. I am working with my co-advisers, Joan Garfield and Michelle Everson.

For the purpose of this instrument, statistical literacy is defined as being able to read, understand, and communicate statistical information. The purpose of the BLIS Assessment is to determine what statistical literacy skills students have in an introductory statistics course, at the postsecondary level. Because it is to be used as an assessment in a more modern version of the introductory course, this instrument includes items on simulation-based methods as well as some of the more traditional topics.

In light of changes to the introductory statistics course, namely the move to include simulation-based methods, a new assessment of statistical literacy is needed. A new statistical literacy assessment may be used for a variety of purposes. For example, the assessment could be used as a pretest to determine what statistical literacy skills students have prior to taking an introductory statistics course at the postsecondary level. Statistics instructors could also use the assessment in their classes to see which statistical literacy topics students understand and which topics they do not understand. Further, researchers can use the assessment to determine how much statistical literacy students gain in an introductory statistics course that incorporates simulation-based methods in the curriculum, or to determine if the amount of statistical literacy gained in introductory statistics courses taught with varying teaching methods or curricula is different.

I am requesting your help at this time to administer the instrument to your students. The data gathered will be used to evaluate the validity and reliability of the BLIS Assessment.

If you agree to do this, I will provide summary data to you to describe the statistical literacy of your students as well as how they compare to students at other institutions. You can choose to administer the BLIS Assessment to your students either inside or

outside of the classroom. You may want to provide extra credit to your students as an incentive. You will receive your students' scores after they complete the exam. Upon request, I will provide a summary of the results collected from other institutions.

Ideally, I would like you to administer the BLIS Assessment to your students by April 25, 2014. 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.

Please let me know if you are interested by emailing me at sath0166@umn.edu. I will then send you a copy of the BLIS Assessment and additional information for your consideration.

Sincerely,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

G3 Follow-up Letter Sent to Instructors for the Field Test

Dear Professor XXX,

Thank you for being interested in administering the BLIS assessment to your students. The BLIS assessment is an assessment of statistical literacy to be used in a course that includes, to some extent, simulation-based methods. The assessment includes 37 forced-choice items and should take approximately 40 to 50 minutes. I am still putting finishing touches on the assessment, which should be complete by next week. After it is complete, I will send you a link if you would like to look at it before administering it to your students.

Below is detailed information for administering the assessment in your statistics course.

First, I do not expect your students to complete the BLIS assessment in a proctored classroom. However, please instruct your students that they are expected to not work together on the test.

Second, while the BLIS assessment is not a timed test, please allow at least 50 minutes for students to take the test. If students take the test out-of-class, let them know that I expect most students to take about 40 to 50 minutes to complete the test.

Third, at the beginning of the online test, students must provide their first and last names. They will also be asked to provide some additional information that is optional (e.g., gender and ethnicity). They are then asked to enter a unique code that identifies your class which you will be provided with if you agree to administer the assessment to your students.

Finally, in order to gather information on your course characteristics, I would like to ask you to complete a 30-second survey before or after you administer the BLIS assessment. You will be provided with a link to the survey if you agree to administer the assessment to your students. All identifying information from this survey will be removed from your record and only aggregate data will be presented in the sharing of results through publications or talks. Your participation is very much appreciated and is completely voluntary.

Ideally, I would like you to administer the BLIS assessment to your students by April 25, 2014. If this will not be possible, please let me know. After you administer the assessment, I will send you an Excel file of the student scores. I sincerely hope that you will be able to help during this phase of my study.

Please be assured that your participation is very important to provide validity and reliability evidence for the proposed assessment. I really appreciate your help in attaining a broad sample of students.

If you would like to administer the assessment to your students or have any questions, please respond to this email.

Thank you,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

G4 Instructions for Instructors on How to Administer the BLIS-3 Assessment to Their Students

Dear Professor XXX,

Thank you again for your interest in administering the BLIS assessment to your students. Students can take the BLIS assessment online at:

https://umn.qualtrics.com/SE/?SID=SV_e8LViVV9pxpkLkh

Feel free to look at the assessment before administering it to your students. Below is detailed information for administering the assessment in your statistics course.

First, we do not expect your students to complete the BLIS assessment in a proctored classroom. However, please instruct your students that they are expected to not work together on the test.

Second, while the BLIS assessment is not a timed test, please allow at least 50 minutes for students to take the test. If students take the test out-of-class, let them know that we expect most students to take about 40 to 50 minutes to complete the test.

Third, at the beginning of the online test, students must provide their first and last names. They will also be asked to provide some additional information that is optional (e.g., gender and ethnicity). They are then asked to enter a unique code that identifies your class. The code for your class is: INCLUDE CODE
Please provide this code (INCLUDE CODE) to your students along with the online link for the test.

If possible, please administer this test by April 25th. Please send an email when your students have completed the test so I can send you an Excel file of the student scores.

Please be assured that your participation is very important to provide validity evidence for the proposed assessment. I really appreciate your help in attaining a broad sample of students.

In summary:

Link to the test: https://umn.qualtrics.com/SE/?SID=SV_e8LViVV9pxpkLkh

Unique class code: INCLUDE CODE

Thank you,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

G5 Instructor Survey for the Field Test

Dear Professor XXX,

Thank you again for agreeing to administer the BLIS assessment to your students. In order to gather information on your course characteristics, I would like to ask you to complete a 30-second survey before or after you administer the BLIS assessment (the survey link: https://umn.qualtrics.com/SE/?SID=SV_aY7Rd9RKq33MTDn). All identifying information from this survey will be removed from your record and only aggregate data will be presented in the sharing of results through publications or talks. Your participation is very much appreciated and is completely voluntary.

Thank you,

Laura Ziegler
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

Instructions:

Please fill out the following survey for the course you will be administering the BLIS assessment in. If you will be administering the BLIS assessment in different courses, please fill out the survey separately for each course.

1. What is your name?
2. Enter the code provided to you for your class.
3. What is the name of your institution?
4. What country is your institution in? Please select below... (dropdown list included in online version of survey)
5. What type of institution are you administering the assessment at?
 - High school
 - 2-year/technical college
 - 4-year college
 - University
6. What mathematics prerequisites are there for the course that you are administering the assessment in? (select all that apply)
 - No mathematics requirement
 - High school algebra
 - College algebra
 - Calculus
7. Will the assessment be administered in class or out of class?
 - In class
 - Out of class

8. Will the assessment scores be used as a homework, quiz, or exam score for students?
- Yes
 - No
9. Will the assessment scores be used to give extra credit to students?
- Yes
 - No
10. Will you be administering the assessment towards the beginning, middle, or end of the course?
- Beginning
 - Middle
 - End
11. Which of the following simulation-methods have you had students use in the introductory statistics course that you will be administering the BLIS assessment in? (select all that apply)
- Bootstrap confidence intervals
 - Randomization tests
 - Probability simulations such as coin flipping
 - Any other simulations to help understand statistical topics such as sampling distributions and confidence intervals.
12. How valuable do you believe the BLIS assessment will be for statistics educators? Please indicate the value of this assessment by selecting a value from 1 to 4 where 1 is not valuable at all, 2 is a little valuable, 3 is valuable, and 4 is very valuable.

	Not at all valuable 1	A little valuable 2	Valuable 3	Very valuable 4
Instructors				
Educational researchers				

Appendix H

Expert Review Results from the Preliminary Test Blueprint

Table H1

Ratings for the Learning Outcomes in the Preliminary Test Blueprint from the Six Expert Reviewers. Groups of Similar Ratings are Also Included Where Group A is the Highest Rated Group and Group D is the Lowest Rated Group.

Learning outcome number	Learning outcome	Expert review ratings					Rating group
		Not essential for test 1	2	3	Essential for test 4	NA	
1	Understanding of the difference between a sample and population		XX		XXXX		A
2	Understanding that there are some recognizable characteristics of randomly sampled or randomly generated data	X	XX	X	X	X	C
3	Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term	X		XXX	XX		B
4	Understanding of the purpose of random sampling in an observational study			X	XXXX	X	A
5	Understanding that statistics computed from random samples tend to be centered at the parameter		X	XX	XX	X	B
6	Understanding of the purpose of random assignment in an experiment				XXXXXX		A

(continued)

Table H1 (continued)

Learning outcome number	Learning outcome	Expert review ratings					Rating group
		Not essential for test 1	2	3	Essential for test 4	NA	
7	Ability to determine what type of study was conducted	X			XXXXX		A
8	Ability to determine if a variable is quantitative or categorical		XX	XX	XX		B
9	Ability to determine if a variable is an explanatory variable or a response variable		XX	XXX	X		B
10	Understanding of the difference between a statistic and parameter		X	X	XXXX		A
11	Understanding that statistics vary				XXXX		A
12	Understanding of resistant statistics	XX	XX	XX			C
13	Understanding of the gambler's fallacy	XXXXX	X				D
14	Ability to describe and interpret a dotplot			XX	XXXX		A
15	Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data		X	X	XXXX		A
16	Ability to create an appropriate graph to display quantitative data	X	XX		XXX		B

(continued)

Table H1 (continued)

Learning outcome number	Learning outcome	Expert review ratings					Rating group
		Not essential for test 1	2	3	Essential for test 4	NA	
17	Understanding the importance of creating graphs prior to analyzing data		X	XX	XXX		A
18	Understanding of what an empirical distribution represents			XX	XXX	X	A
19	Understanding that an empirical distribution shows how sample statistics tend to vary			X	XXX	XX	A
20	Understanding of the difference between mean of a sample, mean of a simulated sample, and mean of an empirical distribution	X	XX		X	XX	C
21	Ability to create a bootstrap distribution to estimate a proportion	XXX	XXX				D
22	Ability to create a bootstrap distribution to estimate a difference in two proportions	XXX	XXX				D
23	Understanding that simulated statistics in the tails of a bootstrap distribution are not plausible estimates of a population parameter	XX	X	X	XX		C
24	Understanding that a bootstrap distribution tends to be centered at the sample statistic	XX	X	XX	X		C
25	Ability to create a randomization distribution to test the difference between two groups	X	XX		XXX		B

(continued)

Table H1 (continued)

Learning outcome number	Learning outcome	Expert review ratings					Rating group
		Not essential for test 1	2	3	Essential for test 4	NA	
26	Understanding that simulated statistics in the tails of a randomization distribution are evidence against the null hypothesis				XXXXX	X	A
27	Understanding that a randomization distribution tends to be centered at the hypothesized null value			XX	XXX	X	A
28	Ability to interpret a probability in the context of the data	X			XXXXX		A
29	Ability to interpret a percent in the context of the data		X		XXXXX		A
30	Ability to interpret a mean in the context of the data		X		XXXXX		A
31	Understand how a mean is affected by skewness or outliers		XX	X	XXX		A
32	Ability to interpret a standard deviation in the context of the data			X	XXXXX		A
33	Understanding of the properties of standard deviation	X	X	XXX	X		B
34	Ability to estimate a standard deviation from a sample	X	XX	X		XX	D

(continued)

Table H1 (continued)

Learning outcome number	Learning outcome	Expert review ratings					Rating group
		Not essential for test 1	2	3	Essential for test 4	NA	
35	Ability to estimate a standard error from an empirical distribution		XXX	X		XX	D
36	Understanding of how sample size affects the standard error			X	XXXXX		A
37	Ability to interpret a margin of error		XX		XXXX		A
38	Understanding of the properties of a bootstrap interval	XXX	X	X	X		D
39	Understanding that a bootstrap interval provides plausible values of the population parameter	XX	X		XXX		B
40	Understanding of the purpose of a bootstrap interval	XX	X	X	XX		C
41	Understanding of what statistic should be computed to create a bootstrap interval	XXX	XX		X		D
42	Understanding of how the confidence level affects the width of a bootstrap interval	X	X	XX	X	X	C
43	Understanding of the logic of a significance test when the null hypothesis is rejected				XXXXXX		A

(continued)

Table H1 (continued)

Learning outcome number	Learning outcome	Expert review ratings					Rating group
		Not essential for test 1	2	3	Essential for test 4	NA	
44	Understanding of the purpose of a hypothesis test		X		XXXXX		A
45	Ability to compute a p-value using a randomization distribution		X	X	XXXX		A
46	Ability to describe a model for a bootstrap interval (outcomes, probabilities, with or without replacement)	XX	X	X	X	X	C
47	Ability to describe a model for a randomization test (outcomes, probabilities, with or without replacement)		X	XX	X	XX	C
48	Ability to determine a null and alternative hypothesis statement based on a research question		X	X	XXXX		A
49	Understanding of what a significance level is		XX	X	XXX		A
50	Understanding of how a significance level is used	X	X	X	XXX		B
51	Ability to determine statistical significance based on a p-value	X			XXXXX		A

(continued)

Table H1 (continued)

Learning outcome number	Learning outcome	Expert review ratings					Rating group
		Not essential for test 1	2	3	Essential for test 4	NA	
52	Understanding that an experimental design with random assignment supports causal inference				XXXXXX		A
53	Understanding of the factors that allow a sample of data to be generalized to the population				XXXXXX		A
54	Understanding that correlation does not imply causation		X	X	XXXX		A

Table H2

Comments from Reviewers on Specific Learning Outcomes and Changes Made to the Preliminary Test Blueprint

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
1	Understanding of the difference between a sample and population			
2	Understanding that there are some recognizable characteristics of randomly sampled or randomly generated data	5	Randomness: I don't know what you mean by "recognizable characteristics" of a random sample; I do understand that there are long-run patterns in repeated sampling.	Deleted due to low ratings and reviewer comment
3	Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term			
4	Understanding of the purpose of random sampling in an observational study	5	Random samples: In my experience it makes sense to categorization types of studies as sample surveys (random sampling), experiments (random assignment) and observational studies (no randomization). Talking about randomization in an observational study confuses the issue.	Deleted due to overlap with learning outcome 53, this learning outcome had lower ratings than learning outcome 53
5	Understanding that statistics computed from random samples tend to be centered at the parameter			

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
6	Understanding of the purpose of random assignment in an experiment			Deleted due to overlap with learning outcome 52
7	Ability to determine what type of study was conducted	1	I think what you want is for students to pick out whether or not the study used random assignment, not the type of study (e.g., cross-over design with repeated measures)	
		5	Observational studies and experiments: This should be a three-way split, as stated above.	
8	Ability to determine if a variable is quantitative or categorical			
9	Ability to determine if a variable is an explanatory variable or a response variable			
10	Understanding of the difference between a statistic and parameter	5	Statistics and parameters: "Understand that statistics vary in repeated sampling from the same population."	
11	Understanding that statistics vary			Changed to "Understanding that statistics vary from sample to sample"

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
12	Understanding of resistant statistics			Deleted due to low ratings
13	Understanding of the gambler's fallacy			Deleted due to low ratings
14	Ability to describe and interpret a dotplot	1	Why just dotplot?	
		5	Dot plots do not deserve a category by themselves. The idea here should be that data needs to be explored by both numerical and graphical techniques before it can be properly analyzed. If dot plots are singled out, why not bar graphs, box plots, histograms, scatterplots etc.	
15	Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data			
16	Ability to create an appropriate graph to display quantitative data	1	I am torn on this. In general the public does not have access to the raw data so the reading of pre-existing graphs is more important.	Deleted due to reviewer comment and inability to create a selected-choice type of item and access software

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
17	Understanding the importance of creating graphs prior to analyzing data			
18	Understanding of what an empirical distribution represents	5	I don't know what you mean by "empirical distributions." It seems they could be distributions of sample data, simulated sampling distributions of statistics, simulated probability distributions or something else.	Changed to "Understanding of what an empirical sampling distribution represents"
		6	Not sure what you mean here. Are you using empirical distribution to mean simulated sampling distribution? If so, I would rate this a 4	
19	Understanding that an empirical distribution shows how sample statistics tend to vary	1	Is "empirical" what you want here? That just means constructed from the data. How about empirical sampling distribution?	Changed to "Understanding that an empirical sampling distribution shows how sample statistics tend to vary"
		3	By "Empirical Distribution" do you mean empirical sampling distribution? I'm guessing so by the context, but I think of any simulated distribution as an empirical distribution, so might be good to clarify.	
		6	Same comment as above.	

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
20	Understanding of the difference between mean of a sample, mean of a simulated sample, and mean of an empirical distribution	1	Not of themselves, but in interpreting a mean from a distribution, the context would have to be considered	Deleted due to low ratings
21	Ability to create a bootstrap distribution to estimate a proportion	1	Creating distributions doesn't seem as literacy oriented to me	Deleted due to low ratings
		5	This is too much for an intro stat course that covers most of the more traditional topics.	
22	Ability to create a bootstrap distribution to estimate a difference in two proportions			Deleted due to low ratings
23	Understanding that simulated statistics in the tails of a bootstrap distribution are not plausible estimates of a population parameter			Changed to "Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter"

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
24	Understanding that a bootstrap distribution tends to be centered at the sample statistic			Changed to “Understanding that a confidence interval is centered at the sample statistic”
25	Ability to create a randomization distribution to test the difference between two groups	5	Randomization distributions: “Ability to create a randomization distribution to test the difference between two groups in a randomized experiment.”	Deleted due to inability to create a selected-choice type of item and access software
26	Understanding that simulated statistics in the tails of a randomization distribution are evidence against the null hypothesis	3	I think you mean “observed statistics” or “sample statistics” rather than simulated statistics.	Changed to “Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis”
		5	In the experimental context, the randomization distribution seems clear. But, what do you mean by “randomization distribution” in general. If I’m interested in testing hypotheses on a single population proportion, are you thinking of a simulated sampling distribution based on a binomial model (technically, not a randomization distribution). Such distributions are centered wherever the investigator chooses to center them. The bottom two boxes here are too vague.	

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
27	Understanding that a randomization distribution tends to be centered at the hypothesized null value			
28	Ability to interpret a probability in the context of the data			Deleted due to overlap with learning outcome 29, percents were emphasized more than probabilities in textbooks when referring information presented in the media
29	Ability to interpret a percent in the context of the data			
30	Ability to interpret a mean in the context of the data			
31	Understand how a mean is affected by skewness or outliers			
32	Ability to interpret a standard deviation in the context of the data			

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
33	Understanding of the properties of standard deviation	3	Not sure what you mean by “properties of standard deviation”	
34	Ability to estimate a standard deviation from a sample	2	THIS STATEMENT IS NOT CLEAR: Do you mean apply the formula for the standard deviation to a set of data? Give a “guesstimate”? Use the sample standard deviation to stimate the population standard deviation.	Deleted due to low ratings
		3	“Ability to estimate a standard deviation from a sample” – do you mean compute a sample standard deviation? Or estimate it by looking at a picture? If the former, I would explicitly say that.	
		4	I was unclear on these two (refering also to estimate st error), We don’t really talk about a rule of thumb like “if the distribution is approximately normal” than range/6 is approximately equal to the standard deviation; we do talk about/expect students to understand that the standard deviation is approximately equal to the average deviation of values from the mean.	
		5	Standard deviation: What do you mean by “estimate a standard deviation from a sample?” Do you simply mean calculate the sample standard deviation as an estimate of a population standard deviation? It is not clear what you expect students to know about standard deviation.	

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
35	Ability to estimate a standard error from an empirical distribution	2	Again, not clear what you mean by “estimate”. Calculate the standard deviation of the empirical distribution of a sample statistic? Look at a dotplot of sample statistics and guess the size of the standard deviation?	Deleted due to low ratings
		4	Unclear on this one	
		5	Standard errors: I don’t understand the first box. Are you intending this to cover approximating a standard error through simulations of a sampling distribution for a statistic (such as a proportion)?	
36	Understanding of how sample size affects the standard error			
37	Ability to interpret a margin of error	5	Margin of errors: For what?	Deleted due to learning outcome measuring statistical reasoning than statistical literacy
38	Understanding of the properties of a bootstrap interval	1	What do you mean by “properties?”	Deleted due to low ratings
		3	Not sure what you mean by “properties of a bootstrap interval”	

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
39	Understanding that a bootstrap interval provides plausible values of the population parameter			Changed to “Understanding that a confidence interval provides plausible values of the population parameter”
40	Understanding of the purpose of a bootstrap interval			Deleted due to overlap with learning outcome 39, this learning outcome had lower ratings than learning outcome 39
41	Understanding of what statistic should be computed to create a bootstrap interval			Deleted due to low ratings
42	Understanding of how the confidence level affects the width of a bootstrap interval	1	In general everyone uses 95% so I don’t know that this matters to me	Changed to “Understanding of how the confidence level affects the width of a confidence interval”

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
43	Understanding of the logic of a significance test when the null hypothesis is rejected	3	“Understanding of the logic of a significance test when the null hypothesis is rejected” – I would take out “when the null hypothesis is rejected” – we want students to understand the logic of a significance test regardless of whether the null is actually rejected or not	Changed to “Understanding of the logic of a significance test”
		5	Randomization tests: These should be used to enhance the understanding of inference in general, not just in the case of rejecting a null hypothesis.	
44	Understanding of the purpose of a hypothesis test			
45	Ability to compute a p-value using a randomization distribution	3	I would change “Ability to compute a p-value using a randomization distribution” to “Ability to estimate a p-value using a randomization distribution” to avoid having to use technology but to make sure they get the general idea of looking to see how extreme the observed statistic is	Changed to “Ability to estimate a p-value using a randomization distribution”
46	Ability to describe a model for a bootstrap interval (outcomes, probabilities, with or without replacement)	3	Not sure what you mean by “Ability to describe a model for a bootstrap interval” (or randomization test) – do you mean how to generate a bootstrap or randomization distribution? When I think of “models” I think of something like a regression model.	Deleted due to low ratings and reviewer comment

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
47	Ability to describe a model for a randomization test (outcomes, probabilities, with or without replacement)	1	Especially if this is tied to the assumptions or limitations of the model...what can't we say?	Deleted due to low ratings and reviewer comment
		4	I thought this was a bit too curriculum specific...especially in its wording. With replacement sampling would only be bootstrapping which I see as (potentially) less important.	
48	Ability to determine a null and alternative hypothesis statement based on a research question			
49	Understanding of what a significance level is			Deleted due to overlap with learning outcome 50, learning outcome 50 seemed to be more about statistical literacy than this learning outcome
50	Understanding of how a significance level is used			
51	Ability to determine statistical significance based on a p-value			

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
52	Understanding that an experimental design with random assignment supports causal inference	3	“Understanding that an experimental design with random assignment supports causal inference” – I would rephrase it as “Understanding that only an experimental design with random assignment can support causal inference”	Changed to “Understanding that only an experimental design with random assignment can support causal inference”
		4	Seem somewhat redundant with earlier items (Understanding of the purpose of random sampling in an observational study and Understanding of the purpose of random assignment in an experiment)	
53	Understanding of the factors that allow a sample of data to be generalized to the population	4	Seem somewhat redundant with earlier items (Understanding of the purpose of random sampling in an observational study and Understanding of the purpose of random assignment in an experiment)	

(continued)

Table H2 (continued)

Learning outcome number	Learning outcome	Reviewer	Reviewer comment	Changes
54	Understanding that correlation does not imply causation	3	I would replace “correlation does not imply causation” with “association does not imply causation” to make it more general	Deleted due to overlap with learning outcome 52, this learning outcome had lower ratings than learning outcome 52
		4	I find this mantra to be somewhat overused. The key idea is that observational studies do not lead to cause-effect, but randomized experiments do. This is not just a characteristic of the (arbitrary) choice of the ‘correlation’ (Pearson) as your statistic. To me, then, this is less important.	
		5	Scope of conclusions: If the intention is to cover modeling (of which correlation should be a big part) the last box is far to narrow in scope. You need a whole set of objectives around exploring bivariate data and fitting regression models, along with inference for the slope, which can be done nicely by randomization.	

Appendix I

Expert Review Results from the Preliminary Assessment

Table I1

Ratings for the Items in the Preliminary Assessment from the Six Expert Reviewers. For Each Item, Reviewers were Asked How Much They Agreed or Disagreed with the Following Statement: “The assessment item measures the specified learning outcome.” Groups of Similar Ratings are Also Included Where Group A is the Highest Rated Group and Group D is the Lowest Rated Group.

Item number	Expert review ratings					Rating group
	Strongly disagree 1	Disagree 2	Agree 3	Strongly agree 4	NA	
1	X		XXX	XX		B
2		XX	XX	XX		B
3		X	XXX	X		B
4				XXXXXX		A
5		X		XXXXX		A
6	X	X	X	XXX		B
7	X		XXXX	X		B
8		X	X	XXX	X	A
9			XX	XXX	X	A
10				XXXXX	X ^a	A
11		XXX		XX	X ^a	C
12	XXXX	X		X		D
13	X		XX	XXX		B
14		XX	XX	XX		B
15			XXXX	XX		A
16		X	X	XXX	X ^b	A

(continued)

Table I1 (continued)

Item number	Expert review ratings					Rating group
	Strongly disagree 1	Disagree 2	Agree 3	Strongly agree 4	NA	
17			XXXX	XX		A
18	X	XX	XX		X ^a	C
19		XXX	XX		X ^a	C
20			XXXX	XX		A
21		X	XX	XXXX		A
22			XX	XXXX		A
23				XXXXX	X	A
24			XX	XXX	X	A
25			X	XXXX	X ^b	A
26			XXX	XXX		A
27		X	XX	XX	X	B
28			X	XXXX	X	A
29			X	XXX	XX	C
30			XX	XXXX		A
31	X		XX	XXX		B
32		X		XXXX	X ^a	A
33			XX	XXXX		A
34		X	XX	XXX		A
35			XXX	XXX		A
36			X	XXXXX		A
37		X	XX	XX	X	B

^aReviewer chose both a rating of a 2 and a rating of a 3

^bReviewer chose both a rating of a 3 and a rating of a 4

Table I2

Comments from Reviewers on Specific Items

Item number	Item	Reviewer	Reviewer comment	Changes
1	The Pew Research Center surveyed a nationally representative sample of 1,002 adults in 2013. The sample percent of internet users that have had an email or social networking account compromised was 21%. Identify the sample and population you would like to make inferences about.	3	I would take out the word “sample” from the second sentence (gives it away) and say something like “In the survey, ...” or “For the people in the survey, the percent...”. Specify America ... otherwise population isn’t clear	Item was reworded so the second sentence does not include the word <i>sample</i>
2	Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 10 times and the other student flips a coin 50 times. Which student is more likely to get close to half of their coin flips heads up? Explain why you chose the student you did.	1 4	I think it seems more like understanding sampling variation I like the question, I’m not as big a fan of the learning outcome or that this item assesses it. I think this item (which, again, I like!) seems more about “bigger sample size means less variability, better inference”	Item was reworded

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
3	<p>A manufacturer of frozen pizzas produces hamburger pizzas where the true average weight is 400 grams. To check the quality of the manufacturing process, random samples of 25 pizzas are taken daily and the average weight is recorded. Assuming that nothing is wrong with the manufacturing process, which sequence below is the most plausible for the average weight for five samples?</p> <p>a. 81, 389, 405, 424, 441. b. 336, 362, 377, 387, 400. c. 395, 402, 420, 445, 450. d. Any of the above.</p>	4	<p>Why hamburger pizzas---that seems unnecessarily confusing and odd, how about pepperoni or sausage? This question seems to require the implicit assumption that the underlying variability in the process isn't exorbitantly large which, while not invalidating the correct response, impacts the relatively difference in plausibility between the options given</p>	Type of pizza was changed to sausage instead of hamburger
		5	<p>The outcome is ill posed. The centering (unbiasedness) is true for means and proportions, but not true in general. It does not work for standard deviations, for example. The last line should say "average weights of the five samples." Five is too small a number to see a pattern. Almost anything can happen in such a small sequence when sampling from a highly variable population. In fact, this is counter to question 2, which makes a case for the larger number of trials.</p>	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
3		6	never heard of a hamburger pizza. Delete hamburger? See suggested addition above.: Wording: "To check the quality of the manufacturing process, random samples of 25 pizzas are taken daily and the average weight of the pizzas in the sample is recorded."	
4	<p>Suppose a researcher wanted to determine if aspirin reduces the chance of a heart attack. The researcher studied 500 patients that 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?</p> <p>a. Observational b. Experimental c. Survey d. None of the above</p>	4	Too generic of an example if students have heard of the physician's health study which we frequently use as an example of a randomized experiment? Maybe use a less generic (but still accessible) context. Also, this is completely an unreasonable study to do in 2013 given the 20 year old findings from the physician's health study	All reviewers gave this item a rating of a 4 so changes were not made

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
5	<p>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 = subcompact, 2 = compact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.</p> <p>What type of variable is this?</p> <p>a. categorical b. quantitative c. continuous</p>	4	<p>I guess the numbers are put in the question to make it more misleading? I actually find it makes it too misleading. The numbers act as a proxy for 'car size' which could be treated quantitatively. Also, an astute student may equate quantitative and continuous and realize that a must be the correct answer.</p>	<p>All other reviewers gave this item a rating of a 4 so changes were not made</p>

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
6	The student plans to see if there is a relationship between the number of speeding tickets a student gets in a year and the type of vehicle he or she drives. Identify the response variable in this study.	1	In a relationship question, there does not necessarily need to be a response variable. The answer should be...don't know...or not enough information given	Item was changed to be more about making a prediction rather than asking about a general relationship
	a. college students	4	Don't mean to be overly critical, but I'm trying to think like my students....someone in my class would say "But couldn't drivers who tend to get lots of tickets tend to purchase certain kinds of cars"? Probably not worth worrying about, especially since focus is on last year's speeding tickets---maybe emphasize that (past year tickets) more clearly?	
	b. type of car c. number of speeding tickets d. average number of speeding tickets last year	5	Either of the variables could be the response of interest, depending on whether you want to predict number of speeding tickets from type of vehicle (responses can be categorical) or type of vehicle from number of tickets. Better to write a more detailed question where the response of interest is clear and both variables are quantitative. This question confounds the two issues, as some students will think response variables cannot be categorical.	

(continued)

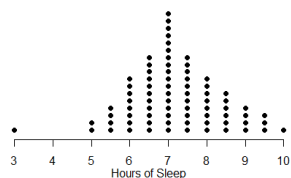
Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
7	CNN conducted a quick vote poll on September 19, 1999 to determine “What proportion of Americans still think that the Miss America pageant is still relevant today?” The poll was conducted on the internet. Here are the results of the poll: Is the Miss American pageant still relevant today? Yes: 1192 votes, No: 4389 votes; Total: 5581 votes. Describe the parameter of interest.	3	Being able to identify the parameter is different than understanding difference between statistic and parameter	Item was changed to ask students to identify both the statistic and parameter
4		While I think the parameter of interest is clear, students have to recognize that they do not need to think about/critique the study design (not enough information is given to suggest this is/isn't a random sample of the population of interest)---perhaps modify to say “random sample of 5581 Americans” and leave it at that?	Item was changed to include a random sample instead of an internet poll	
5		OK, but this is a pretty simple question.		

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
10	The following graph shows a distribution of hours slept last night by a group of college students.	1	It seems that c and d are both ok	Item was taken from the CAOS assessment and asks students to choose the most <i>complete</i> description so the item was not changed



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.

- The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five.
- The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
- Most students seem to be getting enough 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.
- The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours.

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
11	A researcher was interested in determining if a smaller dose of a particular flu vaccine was just as successful as a full dose. A randomized experiment was conducted with 100 participants. Half of the participants received the full dose of the vaccine and the other half received a half does of the vaccine. The number of days the participant got the flu 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 get the flu for the full dose group and half dose group. What step should the research take after the data is collected but before the hypothesis test is conducted?	1	Very vague. What if a student answers..."call a statistician". Or "Enter it into Excel"	Item was changed to specifically mention graphs and why groups should be created
3		there are multiple possible options (calculate summary statistics, for example). Not sure what wrong answers you could give that would be plausible but wrong		
5		The question does not point directly toward graphs. There are a number of possible steps, such as looking for missing data and checking to see that all data values are reasonable simply by looking at a list.		
6		I think tha there are to many ossible good ansers to this question other than what you are looking for here for this outcome...		

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
12	The local Meteorological claims that there is a 70% probability of rain tomorrow. Provide the best interpretation of this statement.	1	This is less about stat and more about meteorology. I was expecting to see something about probability in the long run	A new item was written
	a. Approximately 70% of the city will receive rain within the next 24 hours.			
	b. Historical records show that it has rained on 70% of previous occasions with the same weather conditions.	4	Sorry, I just find this to be too technical of a context to be widely accessible to our students and more about (a) what meteorologists are doing (or think they are doing?) than (b) about a student's understanding of probability	
	c. If we were to repeatedly monitor the weather tomorrow, 70% of the time it will be raining.			
	d. Over the next ten days, it should rain on seven of them.			
		5	This is a poor context for a probability question. Even meteorologists disagree on what such a "probability" means and some think of it as more subjective than objective. Change to a more straightforward context.	
		6	I don't like this item. I have seen several of these interpretations used by meteorologists!	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
13	According to a pet store owner, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean.	1	I think you need to say “in context” or something like that.	Item was changed to ask students to provide context
		4	Why “according to a pet store owner” that seems to open the door up to (a) data came from ‘non-random sample’ and/or (b) data was made-up ☺---maybe according to a national survey of pet owners? Or “Citing a national survey of pet owners, the salesman at a pet store says that ‘the average first-year costs....’”	Item was changed to refer to a national sample
		5	There is not enough context here. The number could simply be a subjective guess, or one particular owner’s value.	

(continued)

Table I2 (continued)

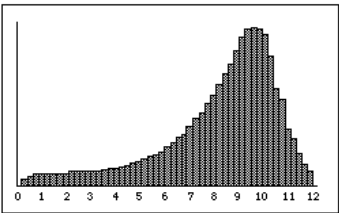
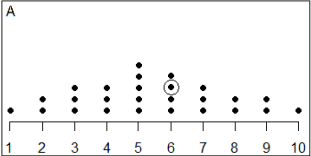
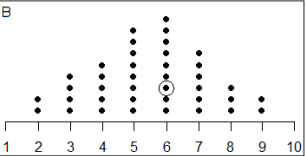
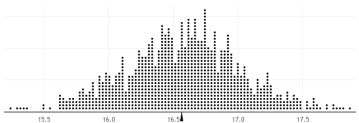
Item number	Item	Reviewer	Reviewer comment	Changes
14	<p>The distribution for a population of measurements is presented below.</p>  <p>A sample of 10 randomly selected values will be taken from the population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?</p> <p>a. 4 to 6 b. 7 to 9 c. 10 to 12</p>	5	None of these options include a symmetric interval around the modal value which, I would guess, is the most common misconception if students don't understand this learning outcome. I don't think adding 9-11 really solves the problem though...maybe...not sure.	<p>Item was changed to include a dotplot instead of a histogram</p> <p>Selected-response options were changed so that the mode was included in one of the intervals</p>
16	<p>A teacher gives a 15 item science test. For each item, a student receives one point for a correct answer; 0 points for no answer; and loses one point for an incorrect answer. Total test scores could range from +15 points to -15 points. The teacher computes the standard deviation of the test scores for the class to be -2.30. What do we know?</p> <p>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.</p>	4 5	<p>There are lots of properties of standard deviation. This doesn't strike me all that important. I think the previous question is good and important. The question itself is good.</p> <p>The question gets at a minor point that has little to do with the deeper understanding of what standard deviation measures. If you are going to use only 37 items, this is a waste of an item.</p>	<p>The item was kept because it measures statistical literacy</p> <p>(continued)</p>

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
17	<p>Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights of a random sample of 3 pebbles each, with the mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.</p> <div style="display: flex; justify-content: space-around;">   </div>	4	<p>I don't know about this one. It is assessing the learning outcome, but my gut reaction is that we can do better in terms of a question---I don't really like the answer choices because it seems like there is only one reasonable answer based on the question set up.</p>	<p>All reviewers gave this item a rating of a 3 or 4 so changes were not made</p>
	<ol style="list-style-type: none"> No, in both Figure A and Figure B, the X represents one pebble that weights 6 grams. Yes, Figure A has a larger range of values than Figure B. Yes, the X in Figure A is the weight for a single pebble, while the X in Figure B represents the average weight of 3 pebbles. 			

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
18	<p>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 graphed by doing the following:</p> <ul style="list-style-type: none"> • 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. <p>Below is the plot of the empirical sampling distribution. (Note: This plot can be used as an estimate of what the sampling distribution would look like.)</p> 	1 3 4 5	<p>Seems quite vague</p> <p>Sampling distribution usually refers to sampling without replacement from the population – this is a bootstrap distribution, not a sampling distribution. Would be confusing to students who actually understand what a sampling distribution is</p> <p>I thought you were eliminating the bootstrap?</p> <p>This is a bootstrapped distribution, not a traditional simulated sampling distribution, which would be confusing to most teachers and students alike. And the question is too open ended. (Even open-ended questions need some guidance as to what you are looking for.)</p>	Item was changed to say “This plot can be used as an estimate of what the sampling distribution would look like.”

What information is obtained from this distribution?

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
19	What values do you believe would NOT be plausible estimates of the population average number of books read? Explain your answer.	1	This seems to need re-wording...maybe given 95% certainty or something. There would likely be a disconnect between plausible (it could happen!) and the fact that the empirical distribution produced values like 17.6, etc. Maybe "Less plausible" rather than NOT	Item was changed to say "less plausible" instead of "not plausible"
		3	even tails are "plausible"... not sure what answer I would give or expect for this question	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
20	The Pew Research Center surveyed 2,076 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% bootstrap interval was 58% to 62%. What is this interval attempting to estimate?	1	This seems more related to describing a parameter.	Item was changed to refer to a “confidence interval” instead of a “bootstrap interval”
		4	I thought you were getting rid of the bootstrap? Too technical, not standard across curricula. This one is measuring the learning outcome as stated here, but doesn’t match the other form you sent?	Item was changed to say that the sample included only Americans
		5	Same concerns about bootstrapping. In addition, I cannot answer the question in any detail because the population from which the sample was selected was never described.	All reviewers gave this item a rating of a 3 or 4 so no additional changes were made
		6	I thought you decided not to do bootstrap intervals??	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
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%)?	4	Not sure how important this is... why is the center of the interval so important relative to other things in the course?	These comments were about the learning outcome, not the item. The learning outcome was reworded to refer to a proportion
	<ul style="list-style-type: none"> 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 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 	5	The learning out come is flawed in that confidence intervals do not have to center on a sample statistic. (Consider exact binomial confidence intervals or confidence intervals for a variance, or even bootstrap intervals for means.) Also, centering of confidence intervals is not the important question. Coverage of the parameter being estimated is the point.	

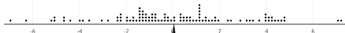
(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
24	Suppose the sample size was doubled from 24 participants to 48 participants and the participants were still split equally into the two groups. Would the standard error change? If so, how? Explain your answer.	4	Again technical language that seems unnecessary...how about just say "variability in the distribution shown above" or range of values or "How would changing the sample size impact the distribution shown above?"	All reviewers gave this item a rating of a 3 or 4, except one reviewer did not provide a rating, so changes were not made
		5	OK, but here have been no previous questions on what a standard error actually is.	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
25	<p>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. 26 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. The average weight loss for the text message group was 2.8 pounds and -2.6 pounds for the control group. Note that the control group had a negative weight loss which means that they actually gained weight. The difference in average weight loss was $(2.8) - (-2.6) = 5.4$.</p> <p>A randomization distribution was graphed by doing the following:</p> <ul style="list-style-type: none"> • From the original sample, the 50 women were re-randomized to the text message group (n=26) and control group (n=24), without replacement. • The mean difference in weight loss was computed for the re-randomized groups and placed on the plot shown below. • This was repeated 99 more times. <p>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.)</p>  <p>(Weight Loss for Text Message Group) – (Weight Loss for Control Group)</p> <p>Why is the randomization distribution centered at 0?</p>	4	The language in this question could be tightened up somewhat. The sentence “The average weight loss for the text message group was 2.8 pounds and -2.6 pounds for the control group.” Reads awkwardly. Also, while “control group” isn’t a real technical term, I’d just say something less technical to be clearer.	Item was changed to specify that the control group did not receive text messages
		5	It should be explained that the control group had no such messages.	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
27	<p>The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?</p> <ul style="list-style-type: none"> a. The circuit is definitely not good and needs to be repaired. b. The electrician decides that the circuit is defective, but it could be good. c. The circuit is definitely good and does not need to be repaired. d. The circuit is most likely good, but it could be defective. 	4	Fairly technical for the proposed learning outcome which doesn't purport to get at technical understanding	All other reviewers gave this item a rating of a 3 or 4, except one reviewer did not provide a rating, so no changes were made

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
29	A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions as a model to make inferences about the two states? Explain your answer.	2	In reference to the question sentence, "I don't believe that this phrasing will be familiar or understandable to most students. Plus, the model is not just the two proportions. The model might be a difference between the two proportions plus sampling variability. Could you remove the words "as a model"?"	Item was changed so it no longer referred to a "model"
		3	not sure what this is getting at... what does "use proportions as a model" mean? Proportions are the statistic and parameter, but not the model. I'm confused...	
		4	I'd be curious to see how many students bring up the fact that the 'random sample' will yield a large number of people who don't participate and what the implications are	
		5	The sampling situation described here (cluster sampling, actually) is too complicated for a question about basic properties of inference. Stick with simple random sampling of voters from each state. (Cluster sampling could work well here, but why confound the issue?)	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
29		6	I think that this item may be overly complicated in that there are sooo many problems introduced.	
31	<p>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?</p> <p>a. A large p-value. b. A small p-value. c. The magnitude of a p-value has no impact on statistical significance.</p>	5	OK, but a very simplistic question. Even getting the correct answer will tell you little about whether the student understands p -values.	All reviewers gave this item a rating of a 3 or 4 except one reviewer so changes were not made

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
32	A clinical trial was conducted to determine if women who have regular mammograms to screen for breast cancer would describe 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 as 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? Explain your answer.	1 4	I think for most students this will work, but for some I can see them saying that it is likely the null is true and then leaving it at that without talking about errors The first sentence is unclear. This seems to be more about accepting the null than it is about type II errors...	Item was not changed because when it was changed to a selected-choice item, it directed students to think about errors
33	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 research question is "Does the dog correctly identify cancer more than half of the time?" The p-value is less than .001. Using a significance level of .05, what conclusion should be made? Explain why you chose to make your conclusion.	5	OK, but so much detail is given in an extreme case that the question almost answers itself with no real thinking required of the student.	All reviewers gave this item a rating of a 3 or 4 so changes were not made

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
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?	4	My students would not be familiar with the terminology in (a) or (c) and (thus) they would not likely choose them.	The wording in selected-response option a was changed to “observational study” and option c was removed
	<ul style="list-style-type: none"> a. Correlational study b. Randomized experiment c. Time Series study d. Survey 	5	The question asks nothing about why a randomized experiment is required here. It is the kind of question for which a student will have the correct answer memorized, with no understanding of why this is the case.	All reviewers, except reviewer 5, gave this item a rating of a 3 or 4 so no additional changes were made

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
36	<p>Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?</p> <div data-bbox="554 464 884 721" style="text-align: center;"> </div> <p>a. Graph A b. Graph B c. Graph C</p>	4	<p>Do you have any sense of pre-course % correct on this item? My guess is it's quite high...begging the question of how necessary it is/whether our course adds any value here.</p>	<p>All reviewers gave this item a rating of a 3 or 4 so changes were not made</p>
		5	<p>Again, the question is not wrong, but too simplistic to gain any real knowledge of the student's understanding of scatterplots. As there is only one question on this topic, this is a wasted opportunity.</p>	

(continued)

Table I2 (continued)

Item number	Item	Reviewer	Reviewer comment	Changes
37	<p>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 model:</p> <p>Predicted Price = $5620 - 440 * \text{Age}$</p> <p>A friend asked him to predict the price of a 5 year old model of this car, using his equation. Which of the following is the most correct response to provide?</p> <ol style="list-style-type: none"> Plot a regression line, find 5 on the horizontal axis, and read off the corresponding value on the y axis. Substitute 5 in the equation and solve for "price". Both of these methods are correct. Neither of these methods is correct. 	5	This is a good algebra question but asks nothing about understanding the statistical ramifications of prediction from regression lines. Again, a wasted opportunity.	All other reviewers gave this item a rating of a 3 or 4, except one reviewer did not provide a rating, so changes were not made

I1 Changes Made to the Preliminary Assessment

1. The Pew Research Center surveyed a nationally representative sample of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. ~~The sample percent of internet users that have had an email or social networking account compromised was 21%.~~ Identify the sample and population about which the Pew Research Center can you would like to make inferences from the survey results and the sample from that population about.
2. Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 50 ~~10~~ times and the other student flips a coin 100 ~~50~~ times. Which student is more likely to get 48% to 52% ~~close to half~~ of their coin flips heads up? Explain why you chose the student you did.
3. A manufacturer of frozen pizzas produces sausage hamburger ~~hamburger~~ pizzas, which have a where the true average weight of is 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 sequence below is the most plausible for the average weights of the for five samples?
 - a. ~~381, 389, 405, 424, 441.~~ 380, 385, 413, 424, 437 (mean = 407.8, sd = 24.67)
 - b. ~~336, 362, 377, 387, 400.~~ 336, 362, 377, 387, 400 (mean = 372.4, sd = 24.64)
 - c. ~~395, 402, 420, 445, 450.~~ 396, 400, 426, 445, 449 (mean = 423.2, sd = 24.63)
 - d. Any of the above.
4. Suppose a researcher wanted to determine if aspirin reduces the chance of a heart attack. The researcher studied 500 patients who ~~that~~ 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
 - d. None of the above

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 = subcompact, 2 = compact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.

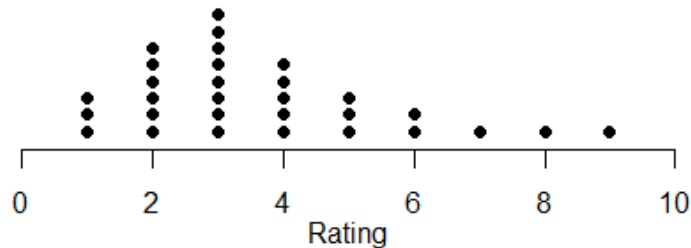
5. What type of variable is this?
 - a. categorical
 - b. quantitative
 - c. continuous

6. The student plans to see if ~~there is a relationship between the number of speeding tickets a student gets in a year and the type of vehicle a student drives~~ he or she 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 ~~car~~
 - c. number of speeding tickets
 - d. average number of speeding tickets last year

7. CNN conducted a quick vote poll with a random sample of 5581 Americans on September 19, 1999 to determine “What proportion of Americans still think that the Miss America pageant is still relevant today?” ~~The poll was conducted on the internet~~. Here are the results of the poll: Is the Miss American pageant still relevant today? Yes: 1192 votes, No: 4389 votes; ~~Total: 5581 votes~~. Identify ~~Describe~~ the statistic and parameter of interest.

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? ~~Researcher B claims that the first study must have been flawed because the mean was not the same in both studies. How would you respond to Researcher B’s statement?~~

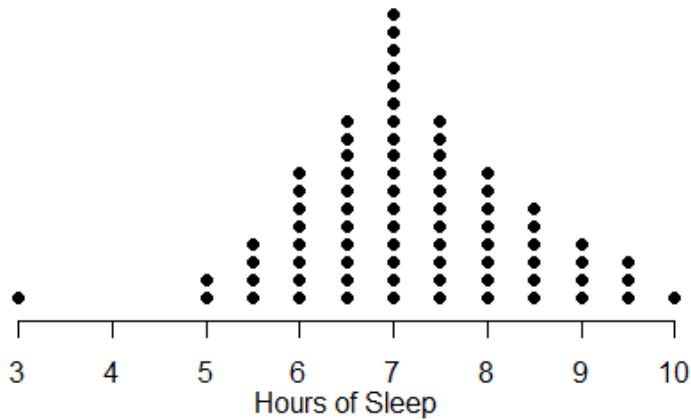
9. One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will ~~aptitude to~~ succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence ~~Aptitude~~ and 10 = Highest Confidence ~~Aptitude~~. The instructor examined the data for men and women separately. Below is the distribution of this variable for the 30 women in the class.



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

- A majority of women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
- The women in the class see themselves as having lower confidence of being able to succeed in ~~aptitude for~~ statistics than the men in the class.
- If you remove the three women with the highest ratings, then the result will show an approximately normal distribution.

10. The following graph shows a distribution of hours slept the previous last 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.

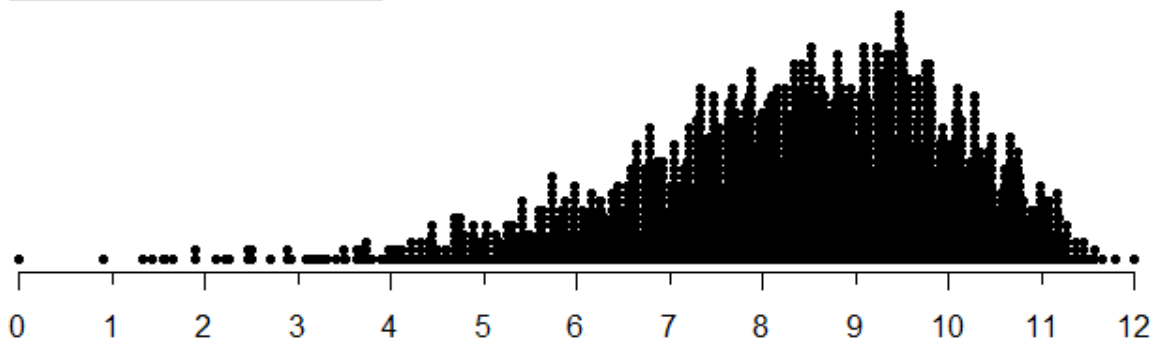
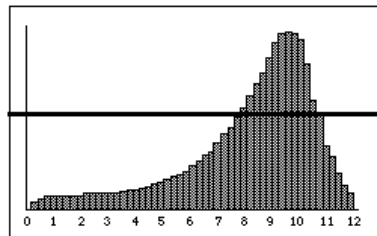
- The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five.
 - The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
 - Most students seem to be getting enough sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.
 - The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours.
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 randomized experiment was conducted with 100 participants. Half of the participants were randomly assigned to received the full dose of the vaccine and the other half received a half dosees of the vaccine. The number of days the participant had get the 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 get the flu symptoms for the full dose group and half dose group. Why should the researcher create and examine graphs of the number of days participants had flu symptoms ~~What step should the research take after the data is collected but before the hypothesis test is conducted?~~

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. Explain what the statistic, .15, means in the context of this report from the National Cancer Institute.

The local Meteorological claims that there is a 70% probability of rain tomorrow. Provide the best interpretation of this statement.

- a. ~~Approximately 70% of the city will receive rain within the next 24 hours.~~
 b. ~~Historical records show that it has rained on 70% of previous occasions with the same weather conditions.~~
 c. ~~If we were to repeatedly monitor the weather tomorrow, 70% of the time it will be raining.~~
 d. ~~Over the next ten days, it should rain on seven of them.~~
13. According to a national survey of pet owners ~~pet store owner~~, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean in context.

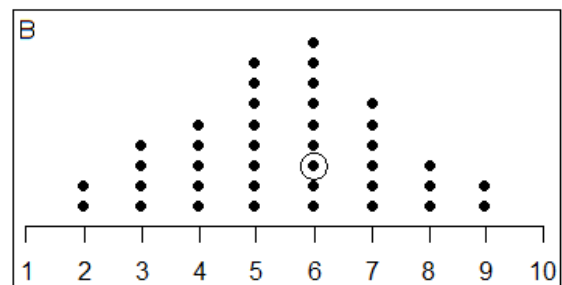
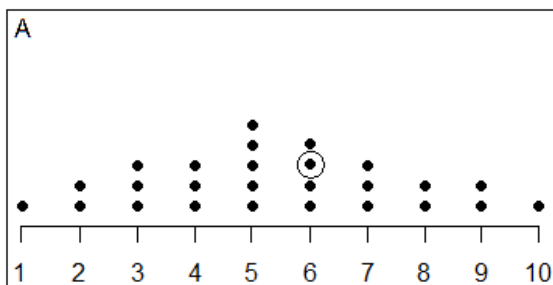
14. The distribution for a population of measurements is presented below.



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 ~~4 to 6~~
 b. 8 to 9 ~~7 to 9~~
 c. 9 to 10 ~~10 to 12~~
 d. 10 to 11

15. ~~Thirty~~ ~~The 30~~ introductory statistics students took a ~~another~~ quiz worth 30 points. The ~~On this quiz, the~~ standard deviation of the ~~quick~~ scores of that quiz was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?
- all of the individual scores are one point apart
 - the difference between the highest and lowest score is 1
 - the difference between the upper and lower quartile is 1
 - a typical distance of a score from the mean is ~~within~~ 1 point ~~of the mean~~
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 computeds the standard deviation of the test scores ~~for the class~~ to be -2.30. What do we know?
- The standard deviation was calculated incorrectly.
 - Most students received negative scores.
 - Most students scored below the mean.
 - None of the above.
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 of a random samples of 3 pebbles each, with all the mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.



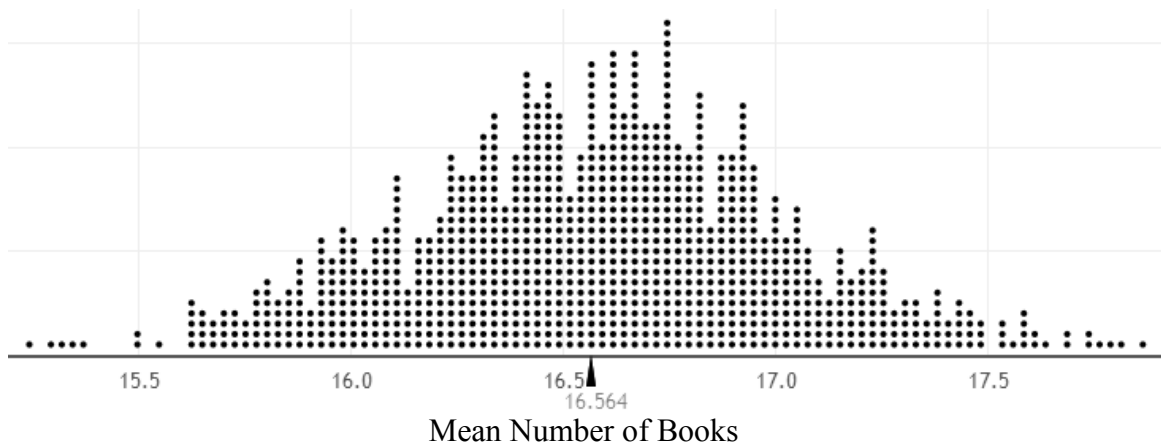
- No, in both Figure A and Figure B, the circled dot X represents one pebble that weights 6 grams.
- Yes, Figure A has a larger range of values than Figure B.
- Yes, the circled dot X in Figure A is the weight for a single pebble, while the circled dot X in Figure B represents the average weight of 3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

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 ~~graphed~~ 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.)



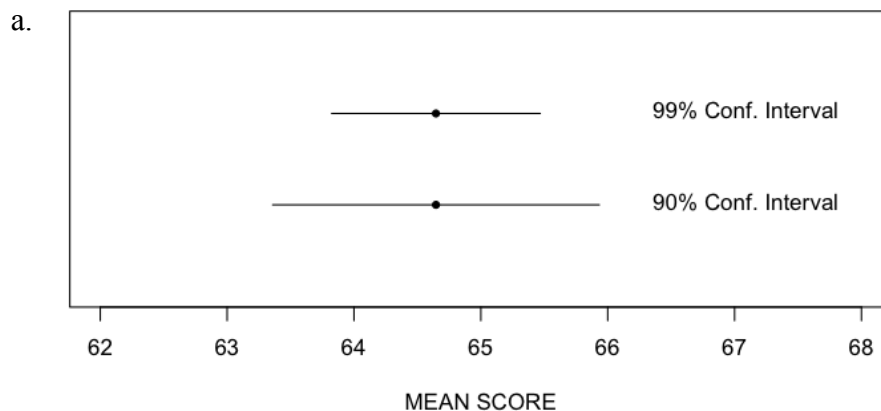
18. What information can be ~~is~~-obtained from this distribution?
19. What values do you believe would **NOT** be LESS plausible estimates of the population average number of books read if you wanted to estimate the population average with 95% confidence? Explain your answer.
20. The Pew Research Center 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 ~~bootstrap~~ interval was 58% to 62%. What is this interval attempting to estimate?

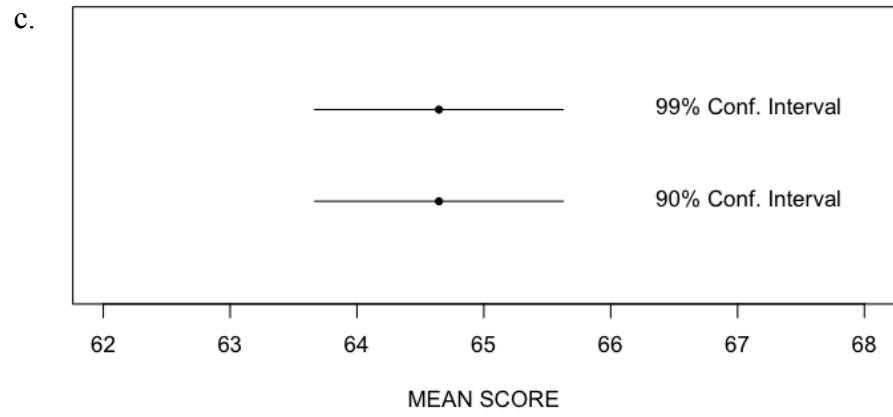
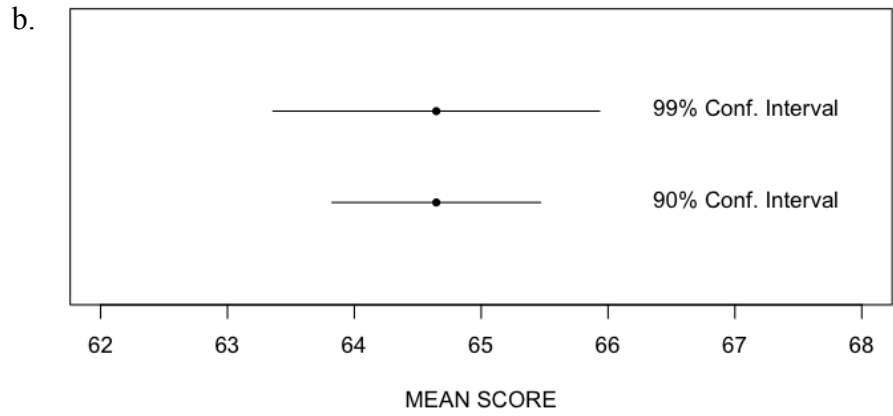
21. In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?
- We know ~~can say~~ that 37% of veterans in the sample have been divorced at least once
 - We know ~~can say~~ that 37% of veterans in the population have been divorced at least once
 - We can say with 95% confidence that 37% of veterans in the sample have been divorced at least once
 - We can say with 95% confidence that 37% of veterans in the population have been divorced at least once
22. This question 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.

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.

Which of the following options would best represent how the two confidence intervals would compare to each other?





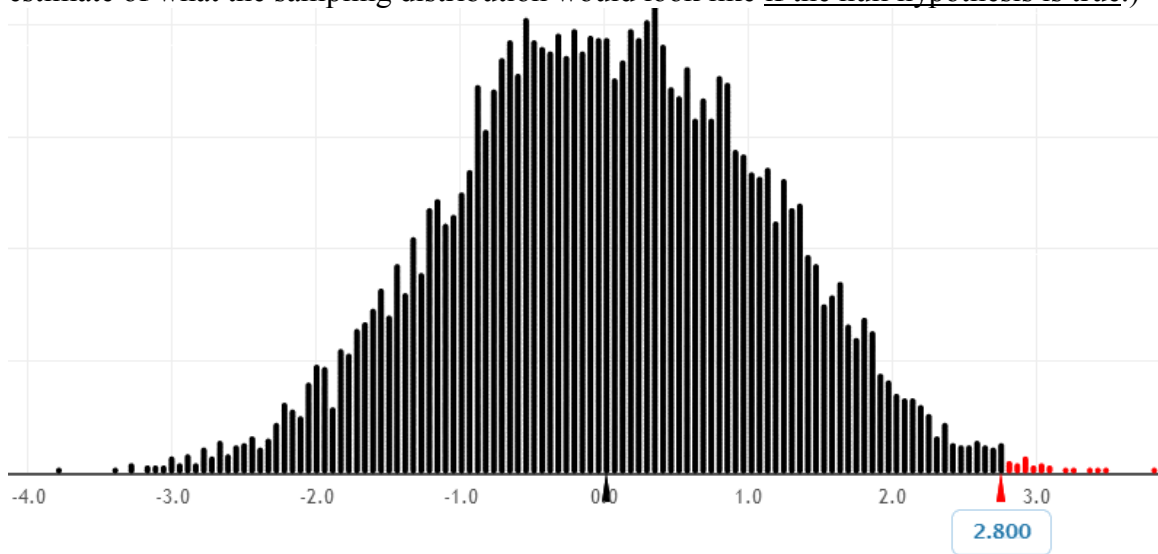
Items 23 and 24 refer to the following situation:

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.

A randomization distribution was ~~produced~~ ~~graphed~~ by doing the following:

- From the original sample, the 24 participants were re-randomized to the nap group (n=12) ~~or~~ ~~and~~ 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)] ~~for the re-randomized groups~~ 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.)



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

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 for the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis? Explain your answer.
24. Suppose the sample size was doubled from 24 participants to 48 participants and the participants were still randomly assigned split equally into the two groups of equal size. How would you expect the standard error to change? If so, how? Explain your answer.

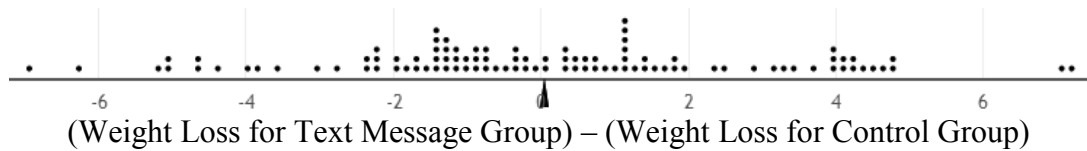
Items 25 and 26 refer to the following situation:

An experiment was conducted with 50 obese women. All women participated in a weight loss program. Twenty-six 26 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 ~~was 2.8 pounds~~ and -2.6 pounds for the control group. Note that the control group had a negative weight loss which means that they actually gained weight. The difference in average weight loss was $(2.8) - (-2.6) = 5.4$.

A randomization distribution was produced ~~graphed~~ by doing the following:

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



25. Why is the randomization distribution centered at 0?

26. Researchers are interested in whether ~~The alternative hypothesis is that the~~ text messages lead to more a higher weight loss than no text messages for women participating in this weight loss program. ~~Therefore a one-tailed (i.e., one-sided) test will be conducted.~~ Compute the approximate p-value for the observed difference in mean weight loss of 5.4 based on the randomization distribution ~~simulated data using the one-tailed test appropriate to the researchers' interest.~~ Explain so someone else can replicate your work how you found his p-value. Show how to find this p-value and explain each step.

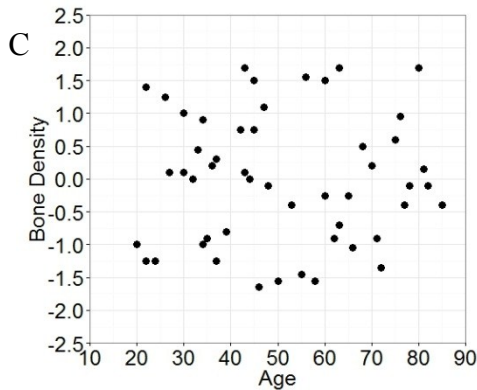
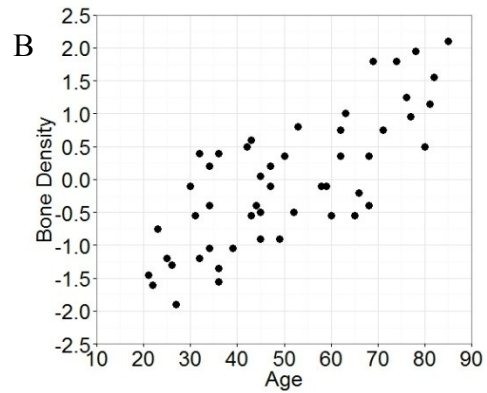
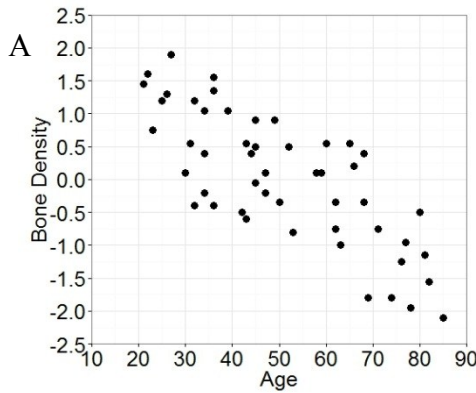
27. The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?
- The circuit is definitely not good and needs to be repaired.
 - The electrician decides that the circuit is defective, but it could be good.
 - The circuit is definitely good and does not need to be repaired.
 - The electrician decides that the circuit is ~~most likely~~ good, but it could be defective.
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~~ 20 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. 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 statistics (70%) as evidence of the effectiveness? Explain your answer.
29. A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions ~~as a model~~ to make inferences about the two states? Explain your answer.
30. 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 days and 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 is difference for males and females with regards to the average number of nights spent in a place not intended for housing?” In order to conduct a hypothesis test to answer this research question, what would the null and alternative hypothesis statements be?

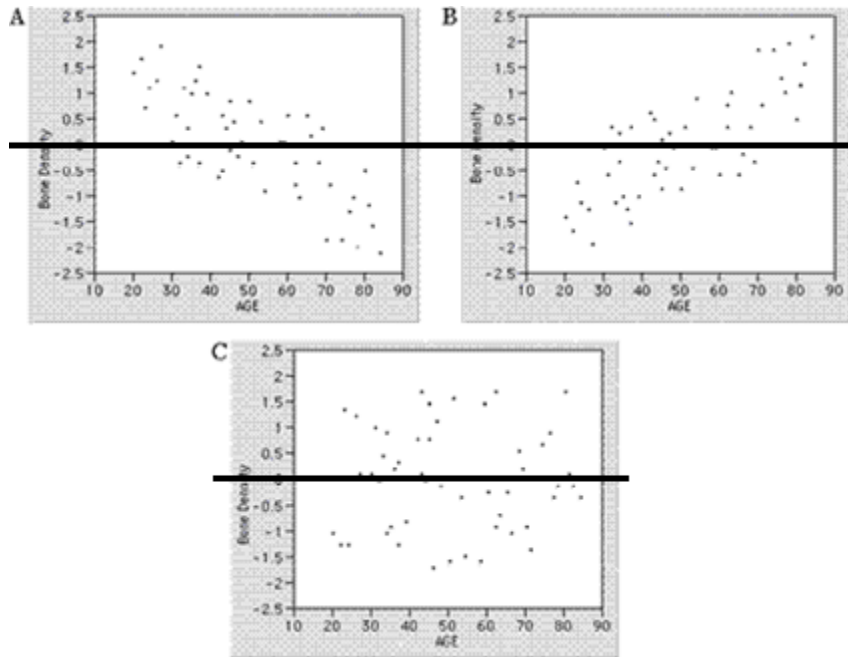
31. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of p -value would she want to obtain?
- A large p -value.
 - A small p -value.
 - The magnitude of a p -value has no impact on statistical significance.
32. A clinical trial was conducted to determine if women who have regular mammograms to screen for breast cancer would decrease ~~describe~~ 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 as 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? Explain your answer.
33. 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 research question is “Does the dog correctly identify cancer more than half of the time?” The p -value is less than .001. Using a significance level of .05, what conclusion should be made? Explain why you chose to make your conclusion.
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?
- Observational ~~Correlational~~ study
 - Randomized experiment
 - ~~Time Series~~ study
 - Survey

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

- a. ~~Although 5,000~~ Five thousand students live in dormitories on campus. A random sample of 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.

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





- Graph A
- Graph B
- Graph C

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 model:

$$\text{Predicted Price} = 5620 - 440 * \text{Age}$$

A friend asked him to use his equation to predict the price of a 5 year-old model of this car, using his equation. Which of the following methods can be used ~~is the most correct response~~ to provide an estimate?

- Plot a regression line, find 5 on the horizontal axis, and read off the corresponding value on the y axis.
- Substitute 5 in the equation and solve for "Predicted Pprice".
- Both of these methods are correct.
- Neither of these methods is correct.

Appendix J

Student Results from the Cognitive Interviews

Table J1

Comments from Students that were Used to Make Changes to Assessment Items to Create the BLIS-2 Assessment

Item number	Item	Student	Student comment	Changes
1	The Pew Research Center surveyed a nationally representative sample 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.	1	This student felt like he was just restating what was stated in the problem. “I am a little bit confused... is it simply asking me to re-iterate that the sample being surveyed is the sample from that population.”	The wording was changed from “The Pew Research Center surveyed a nationally representative sample” to “The Pew Research Center surveyed a nationally representative group”

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
3	A manufacturer of frozen pizzas produces sausage pizzas, which have a true average weight of is 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 sequence below is the most plausible for the average weights of the five samples?	3	This student seemed to think the question was too easy and that there was a trick.	The selected-choice options were changed to include dotplots with 20 sample means in each
		6	This student wanted to choose both selected-response options a and c. He tried to use the empirical rule.	
	a. 380, 385, 413, 424, 437 (mean = 407.8, sd = 24.67)			
	b. 336, 362, 377, 387, 400 (mean = 372.4, sd = 24.64)			
	c. 396, 400, 426, 445, 449 (mean = 423.2, sd = 24.63)			
	d. Any of the above.			

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
4	<p>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?</p> <ol style="list-style-type: none"> Observational Experimental Survey None of the above 	3	This student thought the answer was a “randomized trial” so she chose selected-response option d.	Selected-response option d was deleted
5	<p>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 = subcompact, 2 = compact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.</p> <p>What type of variable is this?</p> <ol style="list-style-type: none"> categorical quantitative continuous 	3	<p>This student did not know what a <i>subcompact</i> was and she thought about that for a while instead of thinking about the question.</p> <p>“I didn’t know what subcompact was until I read compact.”</p>	The order of <i>subcompact</i> and <i>compact</i> was switched

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
7	CNN conducted a quick vote poll with a random sample of 5581 Americans on September 19, 1999 to determine “What proportion of Americans still think that the Miss America pageant is still relevant today?” Here are the results of the poll: Is the Miss American pageant still relevant today? Yes: 1192 votes, No: 4389 votes. Identify the statistic and parameter of interest.	1	“I noticed that the phrasing of the question is a little redundant.”	Redundancy was removed by deleting “Is the Miss American pageant still relevant today?”

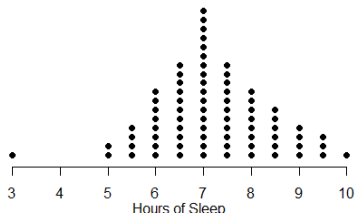
(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
9	<p>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. The instructor examined the data for men and women separately. Below is the distribution of this variable for the 30 women in the class.</p> <p>d the instructor interpret the women's perceptions regarding their success in the class?</p> <ol style="list-style-type: none"> A majority of women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding. The women in the class see themselves as having lower confidence of being able to succeed in statistics than the men in the class. If you remove the three women with the highest ratings, then the result will show an approximately normal distribution. 	1 3	<p>This student was torn between choosing selected-response options a and c.</p> <p>This student talked about the mean and shape and tried to match that with the selected-response options.</p> <p>She was also confused by option b: "Um, I guess I am a little like the men's thing, so, I don't know maybe this is just one of those questions that kind of like, um, where you can't really say or not...so it's kind of a trick."</p>	<p>The question was changed to refer only to students, not men and women</p> <p>The selected-choice options were deleted to change the item into a constructed-response item for the pilot administration of the assessment</p>

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
10	<p>The following graph shows a distribution of hours slept the previous night by a group of college students.</p>  <p>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.</p> <ol style="list-style-type: none"> The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five. The distribution is normal, with a mean of about 7 and a standard deviation of about 1. Most students seem to be getting enough 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. The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours. 	5	<p>“So automatically, like my one I going to be tied between is b and d because they are using a lot of the terms in statistics.”</p> <p>“I’m not like familiar with the range being used.”</p>	<p>Selected-response option c was changed to say “7 hours of sleep” instead of “enough sleep.”</p>
		6	<p>This student seemed a little concerned about using the range.</p> <p>“I will go with d for this one, wait, range is 7 [pause] yeah, d.”</p>	<p>Option d was changed to include the standard deviation instead of the range</p>

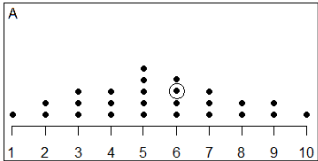
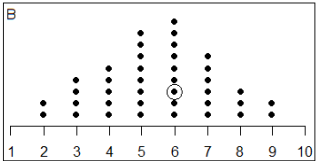
(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
13	According to a national survey of pet owners, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean in context.	6	This student did not like the design of the study. “I feel like this study isn’t completely, [pause] this survey isn’t the best because it is asking pet owners in general but then can it could be cat owners and fish owners and other pets so it should be targeted more to dog owners.”	The item was changed to refer to dog owners, not pet owners

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
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.	1	This student got caught up on selected-response option b.	The selected-choice options were deleted to change the item into a constructed-response item for the pilot administration of the assessment
		3	<p>“Answer b is incorrect cause there is no range for Figure A, oh wait, no, Figure A has a larger range. That isn’t even talking about the point, the circled dot.”</p> <p>This student thought both selected-response options b and c were correct. She had to re-read the question and think about it for a while to realize the question was asking only about the circled dots.</p>	
	 			
	<p>a. No, in both Figure A and Figure B, the circled dot represents one pebble that weights 6 grams.</p> <p>b. Yes, Figure A has a larger range of values than Figure B.</p> <p>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.</p>			

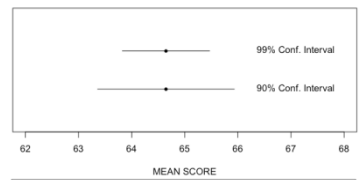
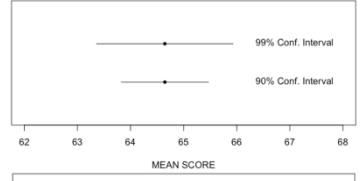
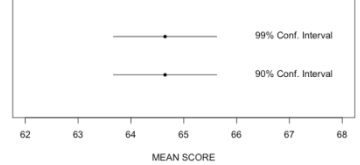
(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
21	<p>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%)?</p> <p>a. We know that 37% of veterans in the sample have been divorced at least once</p> <p>b. We know that 37% of veterans in the population have been divorced at least once</p> <p>c. We can say with 95% confidence that 37% of veterans in the sample have been divorced at least once</p> <p>d. We can say with 95% confidence that 37% of veterans in the population have been divorced at least once</p>	3	This student had to re-read the question a couple of times to figure out the difference between selected-response options a and b, and the difference between options c and d.	The words <i>sample</i> and <i>population</i> in the selected-response options were italicized

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
22	<p>This question 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.</p> <p>Consider a standardized test that has been given to thousands of high school students.</p> <p>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 <i>and</i> a 90% confidence interval for the population mean are constructed using this new sample.</p> <p>Which of the following options would best represent how the two confidence intervals would compare to each other?</p> <p>a. </p> <p>b. </p> <p>c. </p>	5	This student needed to re-read the question a couple of times.	The length of the item was shortened

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
27	<p>The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?</p> <p>a. The circuit is definitely not good and needs to be repaired.</p> <p>b. The electrician decides that the circuit is defective, but it could be good.</p> <p>c. The circuit is definitely good and does not need to be repaired.</p> <p>d. The electrician decides that the circuit is good, but it could be defective.</p>	6	This student appeared to get lost in all the words. He had to re-read parts of it and paused for a while.	The wording of the item was changed to match another version of the item in Ziegler (2012)

(continued)

Table J1 (continued)

Item number	Item	Student	Student comment	Changes
33	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 research question is “Does the dog correctly identify cancer more than half of the time?” The p-value is less than .001. Using a significance level of .05, what conclusion should be made? Explain why you chose to make your conclusion.	1	<p>This student let his own personal beliefs get in the way of answering the question correctly.</p> <p>“The answer to the research question should be no because this is just a bogus experiment and there is nothing to say that the dog is detecting cancer. It could just be because the dog is detecting that particular individual.”</p>	The item was changed so that it said “Assuming the design of the experiment is good” in the question statement.

(continued)

Table J1 (continued)

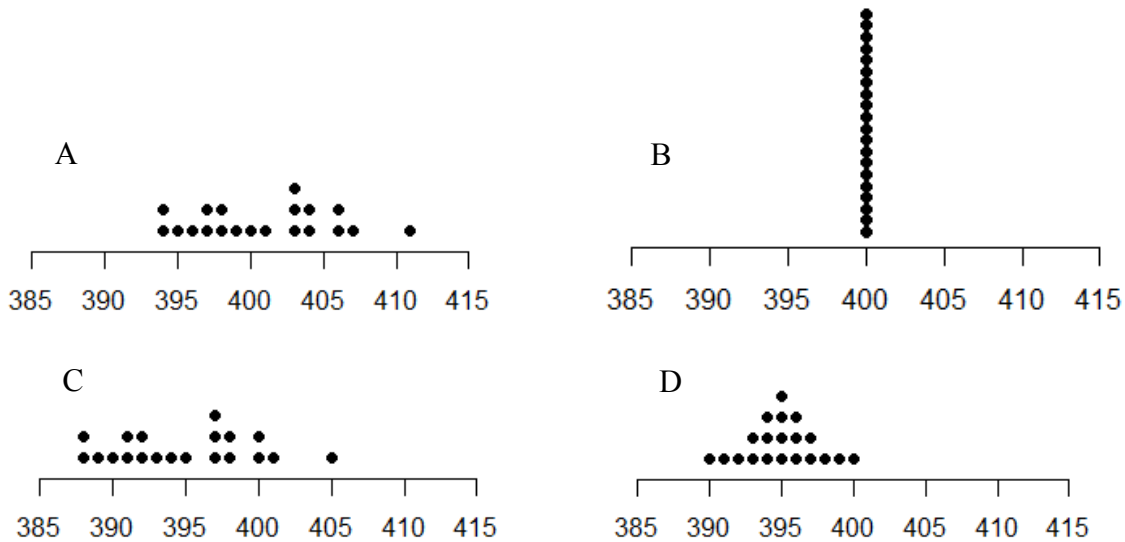
Item number	Item	Student	Student comment	Changes
35	A college official conducted a survey of students currently living in dormitories about their preference for single rooms, double rooms, or multiple (more than two people) rooms in the dormitories on campus. Which of the following does NOT affect the college official's ability to generalize the survey results to all dormitory students?	4	This student chose selected-response option b because he missed the word <i>NOT</i> in the question. He did not even read option c.	The word <i>NOT</i> was underlined and bolded.
	<ul style="list-style-type: none"> 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. 	6	This student first chose selected-response option c and then switched to b. It was clear that he understood the statistical content, but he just missed the word <i>NOT</i> in the question.	

J1 Changes Made to the BLIS-1 Assessment

1. The Pew Research Center surveyed a nationally representative group sample 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.
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? Explain why you chose the student you did.

3. A manufacturer of frozen pizzas produces sausage pizzas, which have a true 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 ~~sequence~~ below is the most plausible for the average weights of 20 ~~the five~~ samples?

- a. ~~380, 385, 413, 424, 437 (mean = 407.8, sd = 24.67)~~
 b. ~~336, 362, 377, 387, 400 (mean = 372.4, sd = 24.64)~~
 c. ~~396, 400, 426, 445, 449 (mean = 423.2, sd = 24.63)~~
 d. ~~Any of the above.~~



- a. Graph A
 b. Graph B
 c. Graph C
 d. Graph D
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
 d. ~~None of the above~~

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 = ~~sub~~compact, 2 = subcompact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and 8 = truck.

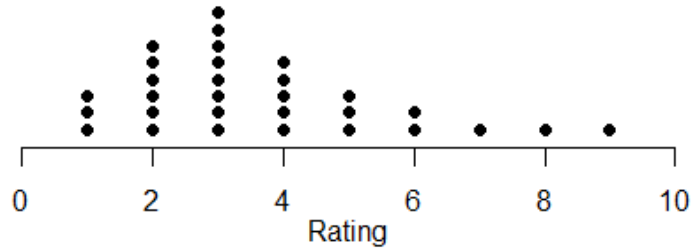
5. What type of variable is this?
 - a. categorical
 - b. quantitative
 - c. continuous

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

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. Here are the results of the poll: Is the Miss American pageant still relevant today? Yes: 1192 votes, No: 4389 votes. Identify the statistic and parameter of interest.

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?

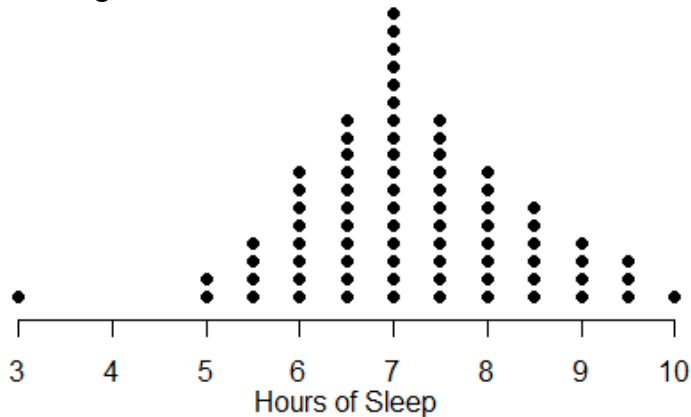
9. One of the items on the student survey for an introductory statistics course was "Rate your confidence that you will succeed in this class on a scale of 1 to 10" where 1 = Lowest Confidence and 10 = Highest Confidence. ~~The instructor examined the data for men and women separately.~~ Below is the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



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

- a. ~~A majority of students women in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.~~
- b. ~~The women in the class see themselves as having lower confidence of being able to succeed in statistics than the men in the class.~~
- c. ~~If you remove the three students women with the highest ratings, then the result will show an approximately normal distribution.~~

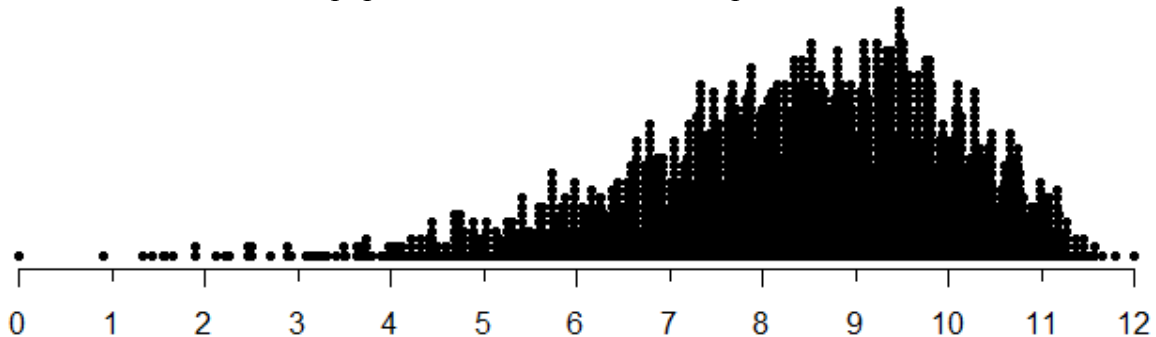
10. The following graph shows a distribution of hours slept the previous night by a group of college students.



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.

- The dots ~~bars~~ go from 3 to 10, increasing in height to 7, then decreasing to 10. The most dots are ~~tallest bar is~~ at 7. There is a gap between three and five.
 - The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
 - Many ~~Most~~ students seem to be getting 7 hours of ~~enough~~ 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.
 - The distribution of hours of sleep is somewhat normal ~~symmetric and bell-shaped~~, with an outlier at 3. The typical amount of sleep is about 7 hours and standard deviation is about 1 ~~overall range is 7~~ hours.
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. Why should the researcher create and examine graphs of the number of days participants had flu symptoms before the hypothesis test is conducted?

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. Explain what the statistic, .15, means in the context of this report from the National Cancer Institute.
13. According to a national survey of dog pet owners, the average first-year costs for owning a large-sized dog is \$1,700. Provide an interpretation of the mean in context.
14. The distribution for a population of measurements is presented below.

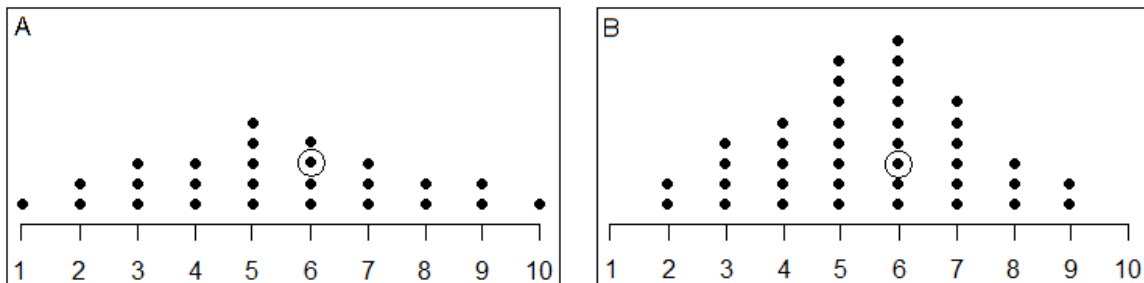


- A sample of 10 randomly selected values will be taken from the population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?
- 6 to 7
 - 8 to 9
 - 9 to 10
 - 10 to 11
15. Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quiz scores was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?
- all of the individual scores are one point apart
 - the difference between the highest and lowest score is 1 point
 - the difference between the upper and lower quartile is 1 point
 - a typical distance of a score from the mean is 1 point

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.

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? Explain your answer.



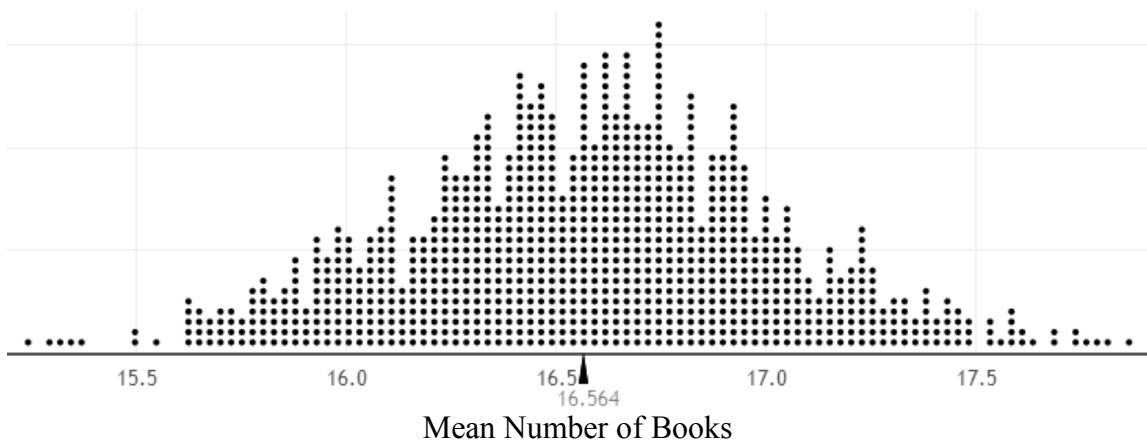
- a. ~~No, in both Figure A and Figure B, the circled dot represents one pebble that weights 6 grams.~~
- b. ~~Yes, Figure A has a larger range of values than Figure B.~~
- c. ~~Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B represents the average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

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.)



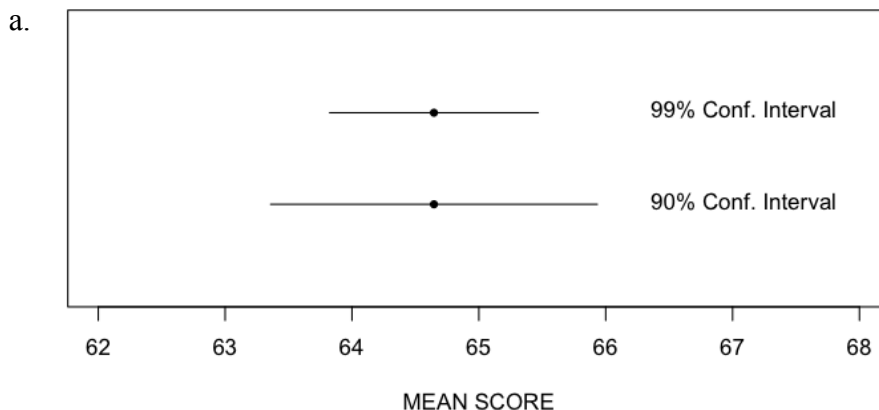
18. What information about the variability from sample to sample can be obtained from this distribution?
19. What values do you believe would be LESS plausible estimates of the population average number of books read if you wanted to estimate the population average with 95% confidence? Explain your answer.
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?

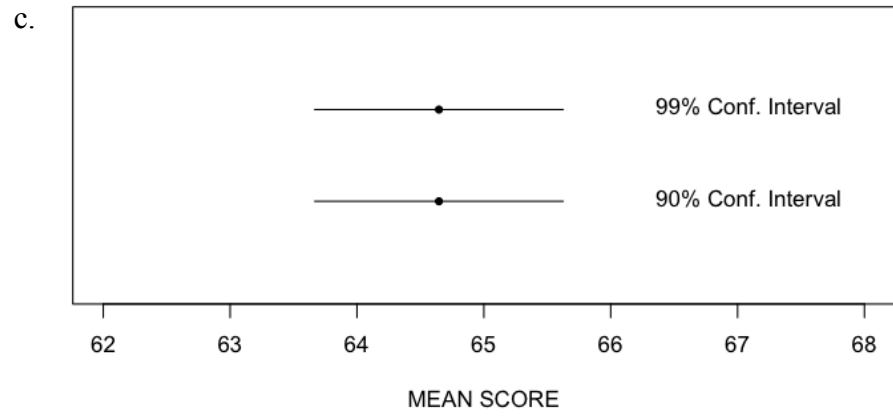
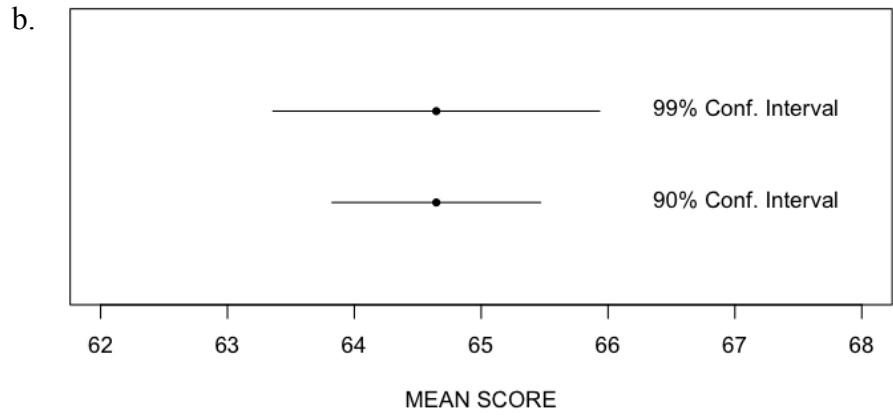
21. In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?
- We know that 37% of veterans in the sample ~~sample~~ have been divorced at least once
 - We know that 37% of veterans in the population ~~population~~ have been divorced at least once
 - We can say with 95% confidence that 37% of veterans in the sample ~~sample~~ have been divorced at least once
 - We can say with 95% confidence that 37% of veterans in the population ~~population~~ have been divorced at least once
22. ~~This question 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.~~

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 ~~following~~ options would best represent how the two confidence intervals would compare to each other?





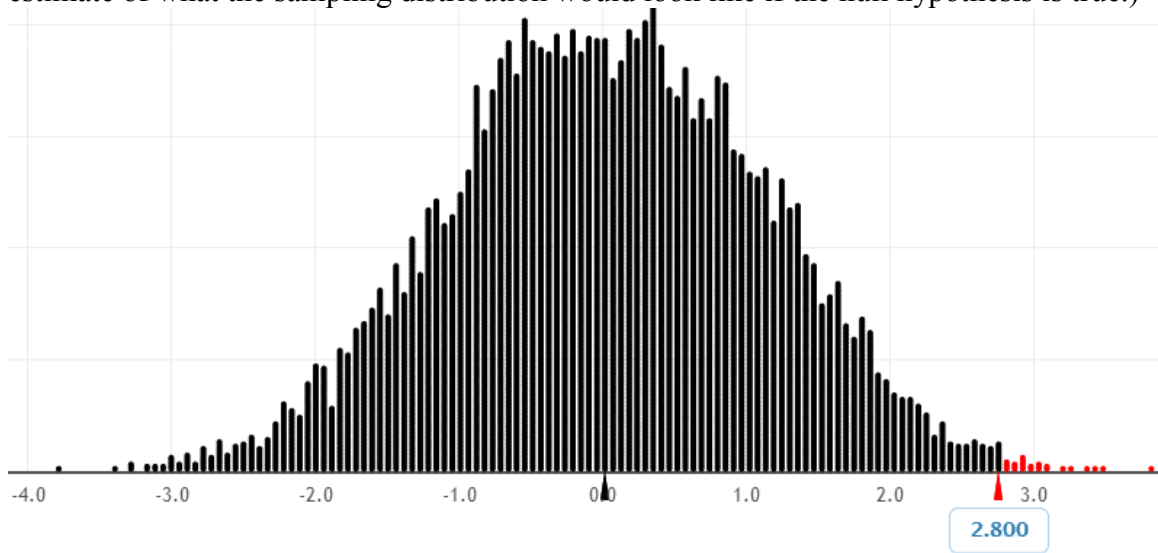
Items 23 and 24 refer to the following situation:

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.

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.)



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

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 for the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis? Explain your answer.

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 to change? Explain your answer.

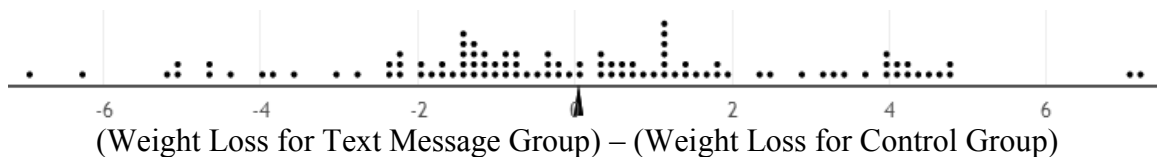
Items 25 and 26 refer to the following situation:

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 weight loss which means that they actually gained weight. The difference in average weight loss was $(2.8) - (-2.6) = 5.4$.

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 [$\text{mean}(\text{text message}) - \text{mean}(\text{control})$] and placed on the plot shown below.
- This was repeated 99 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.)



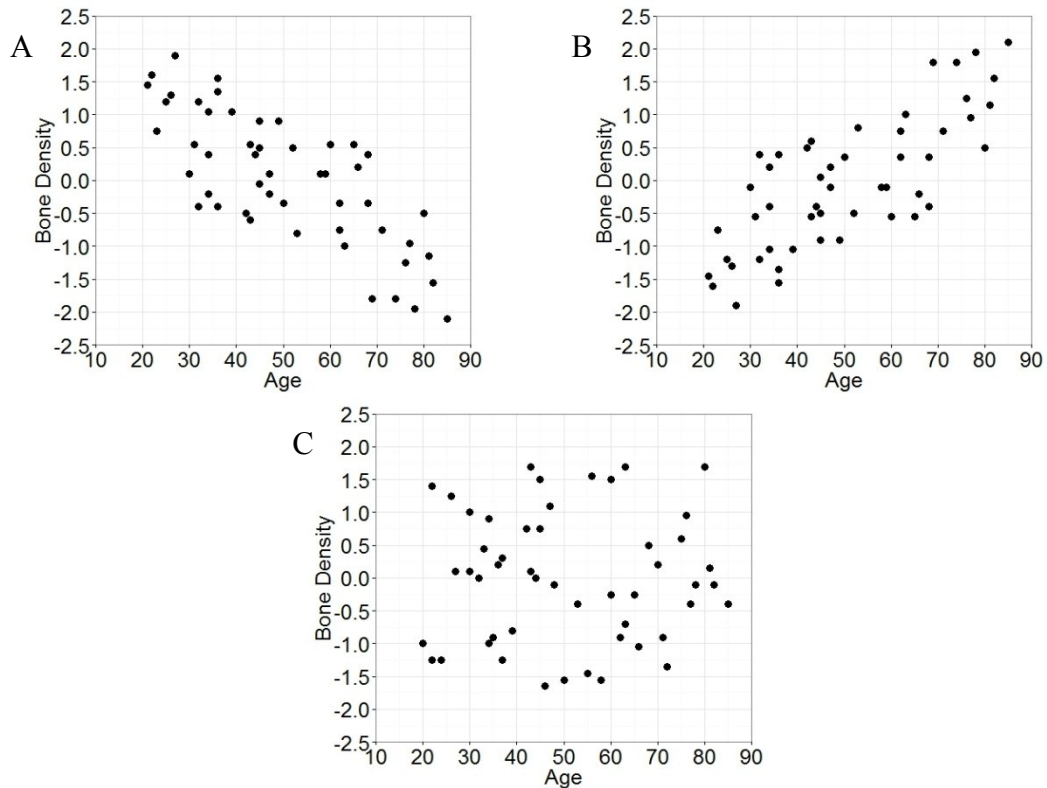
25. Why is the randomization distribution centered at 0?

26. Researchers are interested in whether 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' interest. Explain so someone else can replicate your work how you found his p -value.

27. The following situation models the logic of a hypothesis test. An electrician ~~uses an instrument to test~~ whether or not an electrical circuit is good ~~defective~~. The ~~instrument sometimes fails to detect that a circuit is good and working~~. The null hypothesis is that the circuit is good (~~not defective~~). The alternative hypothesis is that the circuit is not good (~~defective~~). If the electrician performs the test and decides to reject the null hypothesis, ~~w~~ Which of the following statements is true?
- The circuit is definitely not good and needs to be repaired.
 - The ~~electrician decides that the~~ circuit is most likely not good ~~defective~~, but it could be good.
 - The circuit is definitely good and does not need to be repaired.
 - The ~~electrician decides that the~~ circuit is most likely good, but it might not be good ~~could be defective~~.
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. 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 statistics (70%) as evidence of the effectiveness? Explain your answer.
29. A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions to make inferences about the two states? Explain your answer.
30. 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 is difference for males and females with regards to the average number of nights spent in a place not intended for housing?” In order to conduct a hypothesis test to answer this research question, what would the null and alternative hypothesis statements be?

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 large p -value.
 - A small p -value.
 - The magnitude of a p -value has no impact on statistical significance.
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? Explain your answer.
33. 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 research question is that “Does the dog correctly identifies γ cancer more than half of the time?” The p -value is less than .001. Assuming the design of the experiment is good, use Using a significance level of .05 to make a decision. , what conclusion should be made? Explain why you chose to make your decision conclusion.
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?
- Observational study
 - Randomized experiment
 - Survey

35. A college official conducted a survey 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. Which of the following does **NOT** affect the college official's ability to generalize the survey results to all dormitory students?
- Although 5,000 students live in dormitories on campus, only 500 were sent the survey.
 - The survey was sent to only first-year students.
 - Of the 500 students who were sent the survey, only 160 responded.
 - All of the above present a problem for generalizing the results.
36. Studies show that as women grow older they tend to have lower bone density. Which of the following graphs illustrates this point?



- Graph A
- Graph B
- Graph C

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 model:

$$\text{Predicted Price} = 5620 - 440 * \text{Age}$$

A friend asked him to use his equation 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. Plot a regression line, find 5 on the horizontal axis, and read off the corresponding value on the y axis.
- b. Substitute 5 in the equation and solve for "Predicted Price".
- c. Both of these methods are correct.
- d. Neither of these methods is correct.

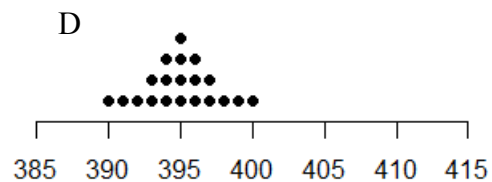
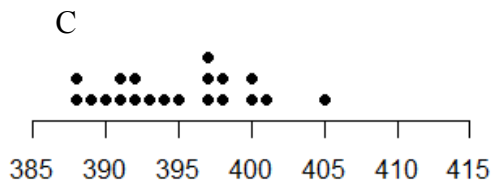
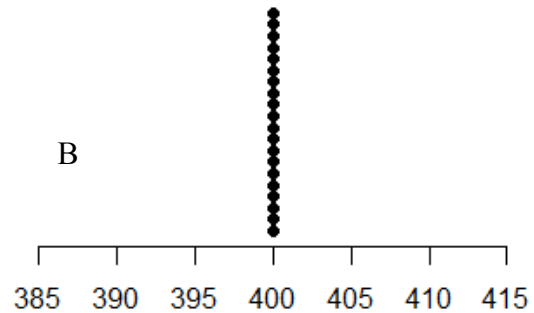
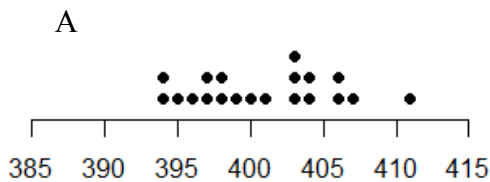
Appendix K

Changes Made to the BLIS-2 Assessment

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.

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? ~~Explain why you chose the student you did.~~
 - 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.

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?



- Graph A
 - Graph B
 - Graph C
 - Graph D
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?
- Observational
 - Experimental
 - Survey

Items 5 and 6 refer to the following situation:

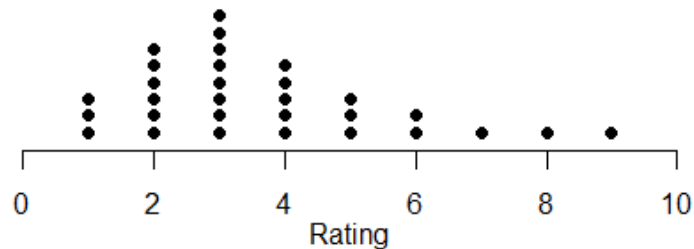
A student is gathering data on the driving experiences of other college students. One of the variables measured is the type of car the student drives. These data are coded using the following method: 1 = compact, 2 = subcompact, 3 = standard size, 4 = full size, 5 = premium, 6 = mini van, 7 = SUV, and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

5. What type of variable is this?
 - a. Categorical
 - b. Quantitative
 - c. Continuous

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

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.

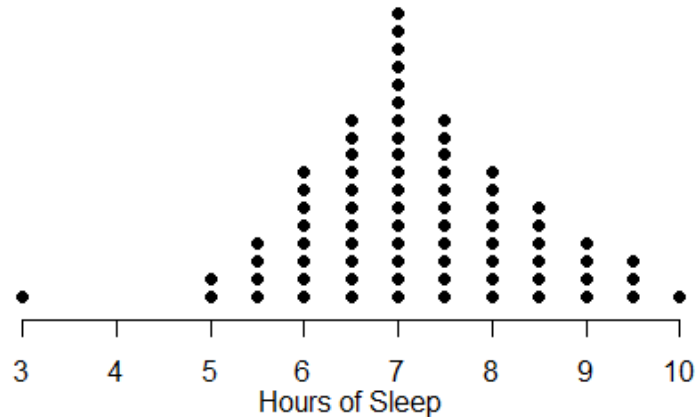
8. In a study, Researcher A took a random sample of 25 college students and found the mean number of times they went out to eat during the last week was 4.1. In another study, Researcher B took a random sample of 25 students from the same college and found the mean number of times they went out to eat 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?
- The sample means varied because they are small samples.
 - The sample means varied because the samples were not representative of all college students.
 - The sample means varied because each sample is a different subset of the population.
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 majority of students in the class do not feel that they will succeed in statistics although a few feel confident about succeeding.
- A majority of students in the class rated their confidence as a 3 although some ratings were higher and some ratings were lower.
- A majority of students will not try to do well in the course because they do not feel that they will succeed in statistics.

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

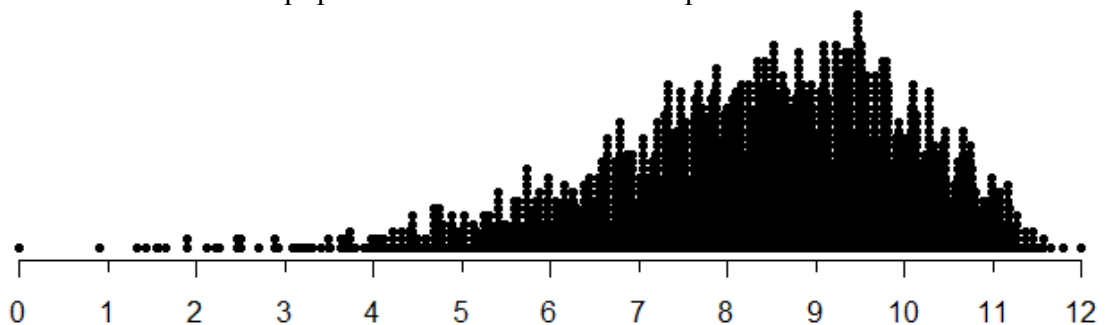


Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.

- 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 distribution is normal, with a mean of about 7 and a standard deviation of about 1.
 - 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.
 - 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.
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 should the researcher should create and examine graphs of the number of days participants had flu symptoms before the hypothesis test is conducted?
- To decide what the null hypothesis and alternative hypothesis should be.
 - To compute the average number of days participants had flu symptoms in order to conduct a hypothesis test.
 - To see if there are recognizable differences in the two groups to decide if a hypothesis test is necessary.

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. ~~Explain~~ What does the statistic, .15, mean in the context of this report from the National Cancer Institute?
- For all men living in the United States, approximately 15% will develop prostate cancer at some point in their lives.
 - 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.
 - In a random sample of 100 men in the United States, 15 men will develop prostate cancer.
 - Both a and b are correct.
13. According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is \$1,700. ~~Provide an~~ Which of the following is the best interpretation of the mean? in context.
- 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.

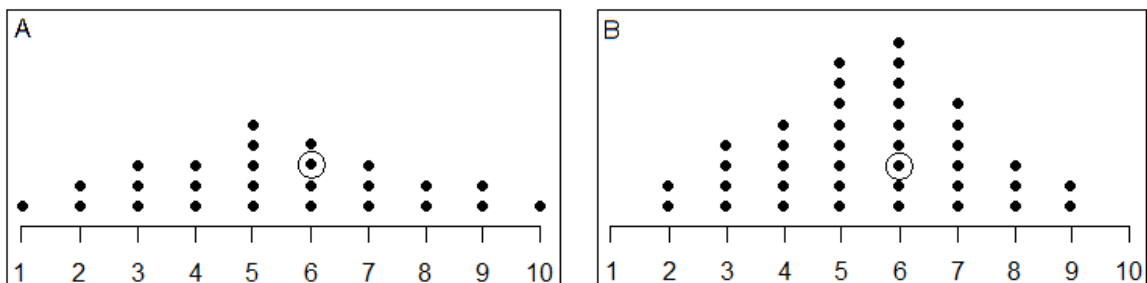
14. The distribution for a population of measurements is presented below.



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?

- 6 to 7
- 8 to 9
- 9 to 10
- 10 to 11

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?
- All of the individual scores are one point apart.
 - The difference between the highest and lowest score is 1 point.
 - The difference between the upper and lower quartile is 1 point.
 - A typical distance of a score from the mean is 1 point.
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?
- The standard deviation was calculated incorrectly.
 - Most students received negative scores.
 - Most students scored below the mean.
 - None of the above.
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. Explain your answer.



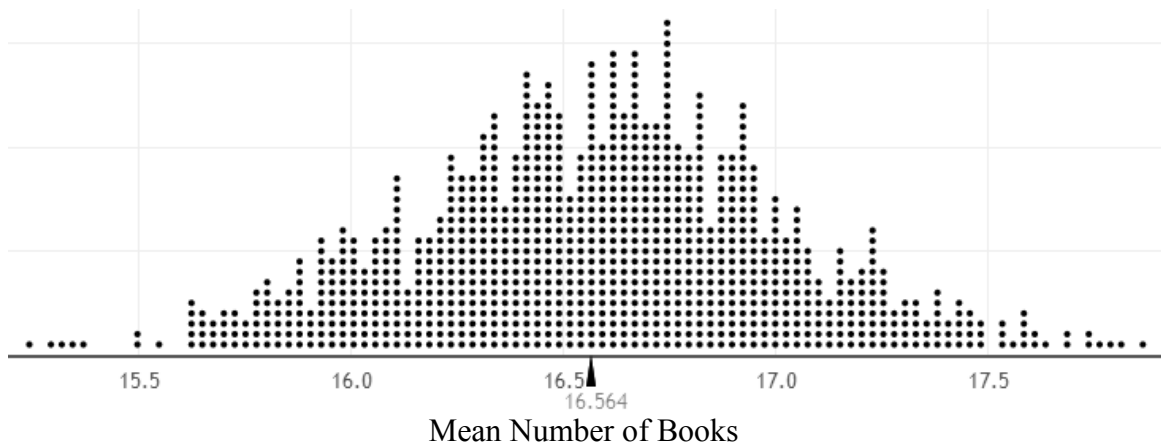
- No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
- Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6.
- Yes, the circled dot in Figure A is the weight for a single pebble, while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

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.)



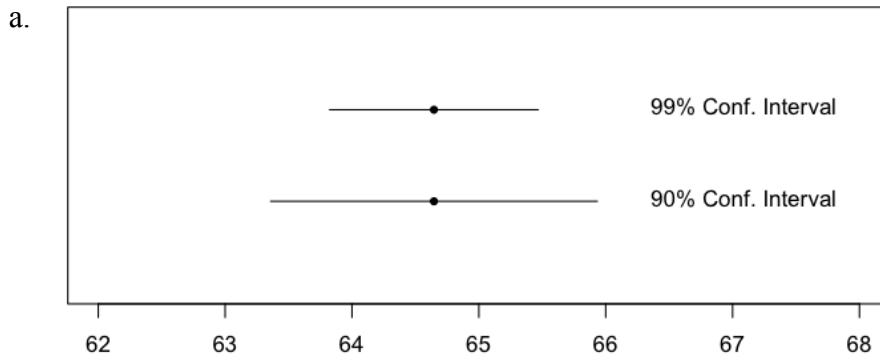
18. Which of the following is the best description of ~~What information about the variability in the empirical sampling distribution from sample to sample can be obtained from this distribution?~~
- The mean number of books that adults read during the year of 2011 was 16.564.
 - The variability in the mean number of books from sample to sample is quite small spanning from approximately 15 to 18.
 - The variability in the number of books from person to person is quite small spanning from approximately 15 to 18.

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? ~~Explain your answer.~~
- Values approximately 17.2 and above because it is unlikely that adults would read that many books.
 - Values below approximately 15.0 and values above approximately 18.0 because there are no dots that are that extreme.
 - Values in the bottom 5% (below approximately 16.0) and values in the top 5% (above approximately 17.0).
 - Values in the bottom 2.5% (below approximately 15.7) and values in the top 2.5% (above approximately 17.4).
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?
- The average number of American adult cell phone users who access the internet on their phones in 2013.
 - The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
 - The percent of all American adult cell phone users who access the internet on their phones in 2013.
 - For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.
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%)?
- We know that 37% of veterans in the *sample* have been divorced at least once.
 - We know that 37% of veterans in the *population* have been divorced at least once.
 - We can say with 95% confidence that 37% of veterans in the *sample* have been divorced at least once.
 - We can say with 95% confidence that 37% of veterans in the *population* have been divorced at least once.

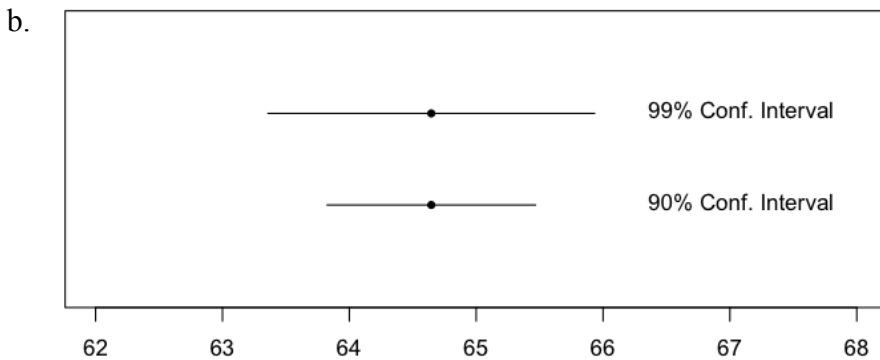
22. Consider a standardized test that has been given to thousands of high school students.

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.

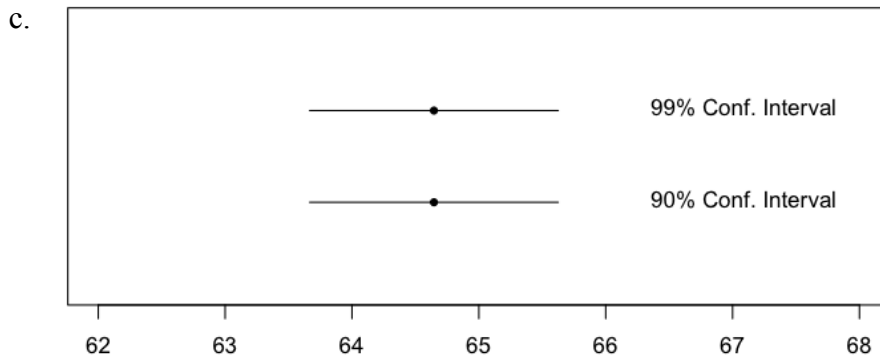
For the following options, a confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval. Which of the options would best represent how the two confidence intervals would compare to each other?



MEAN SCORE



MEAN SCORE



MEAN SCORE

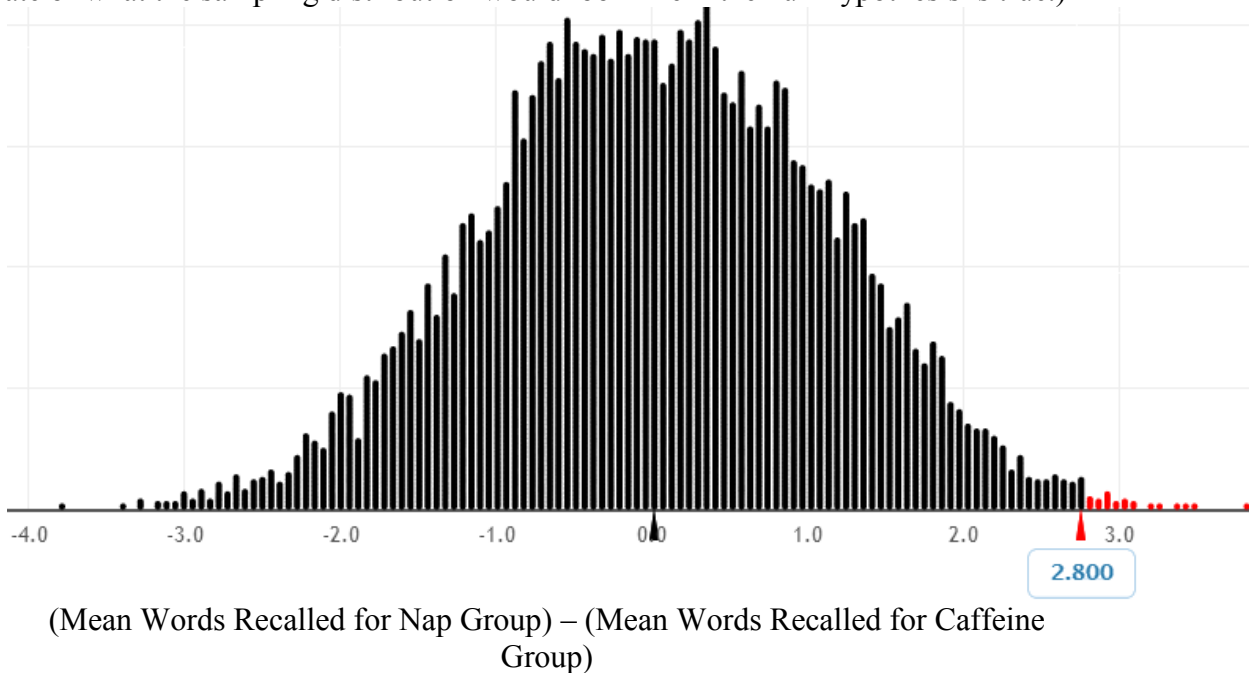
Items 23 and 24 refer to the following situation:

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 [$\text{mean}(\text{nap group}) - \text{mean}(\text{caffeine group})$] 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.)



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? ~~Explain your answer.~~
- 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.
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? ~~Explain your answer.~~
- 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.

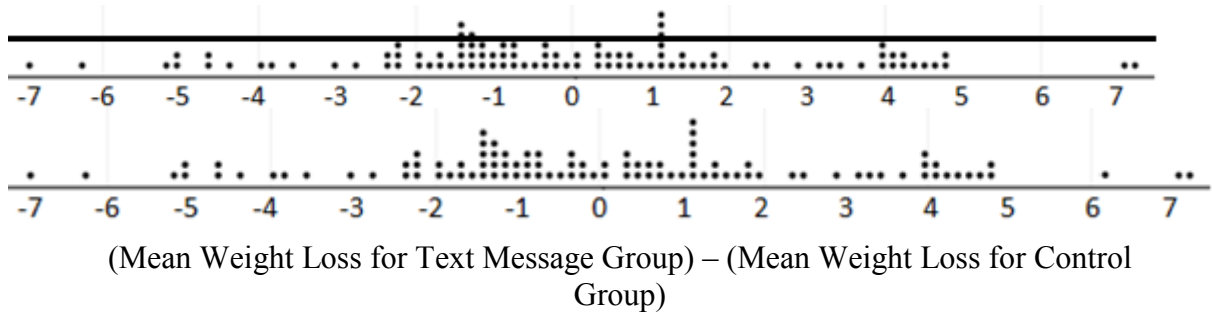
Items 25 and 26 refer to the following situation:

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 [$\text{mean}(\text{text message}) - \text{mean}(\text{control})$] 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.)



25. ~~Explain why~~ Why is the randomization distribution centered at 0?

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

26. Researchers hypothesize that are interested in whether 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. interest. ~~Explain how you found his p -value so someone else can replicate your work.~~
- .03
 - .05
 - .06
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?
- The circuit is definitely not good and needs to be repaired.
 - The circuit is most likely not good, but it could be good.
 - The circuit is definitely good and does not need to be repaired.
 - The circuit is most likely good, but it might not be good.
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? ~~Explain your answer.~~
- The researcher does not need to conduct a hypothesis test because 70% is much larger than 50%.
 - The researcher should conduct a hypothesis test because a hypothesis test is always appropriate.
 - The researcher should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.

29. A news agency wants to know if the proportion of votes for a republican candidate will differ for two states in the next presidential election. For each state, they randomly select 100 residential numbers from a phone book. Each selected residence is then called and asked for (a) the total number of eligible voters in the household and (b) the number of voters that will vote for a republican candidate. The proportion of votes for the republican candidate is computed by adding the number of votes for the republican candidate and dividing by the number of eligible voters for each state. Is it appropriate to use these proportions to make inferences about the two states? Explain your answer.

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?” ~~In order to conduct a hypothesis test to answer this research question, what would the null and alternative hypothesis statements be?~~

29. Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question?
- There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
 - There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
 - There is *no* difference between men and women in terms of the *average number* of nights spent in a place not intended for housing.
 - There is a difference between men and women in terms of the *average number* of nights spent in a place not intended for housing.

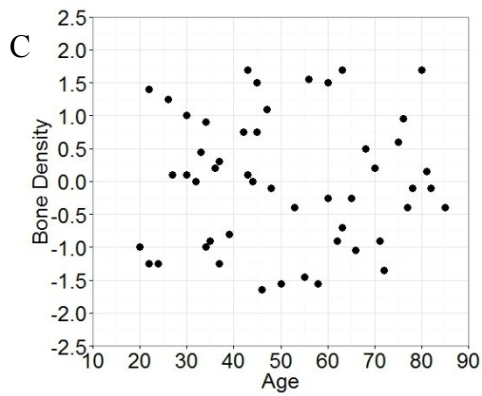
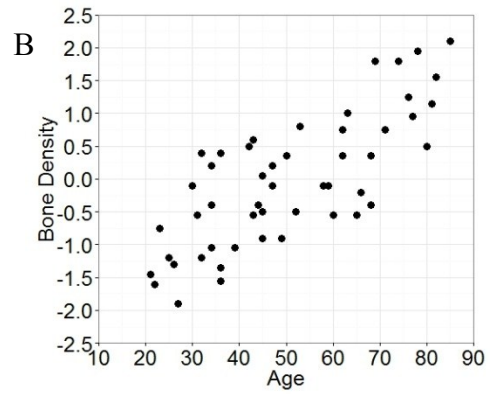
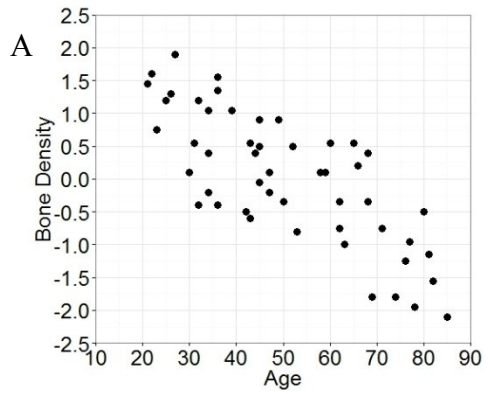
30. Which of the following is a statement of the alternative hypothesis for a statistical test designed to answer the research question?
- There is *no* difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
 - There is a difference between men and women in terms of the *number* of nights spent in a place not intended for housing.
 - There is *no* difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
 - There is a difference between men and women in terms of the *average* number of nights spent in a place not intended for housing.
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 large p -value.
 - A small p -value.
 - The magnitude of a p -value has no impact on statistical significance.
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? Explain your answer.
- Yes. It means you cannot conclude that the alternative hypothesis is true, so the null hypothesis must be true.
 - No. It means you cannot conclude that the null hypothesis is true, so the alternative hypothesis must be true.
 - No. It means that there is not enough evidence to conclude that the null hypothesis is false.
 - No. It means that there is not enough evidence to conclude that the alternative hypothesis is false.

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. ~~Explain why you chose to make your decision.~~
- Reject the null hypothesis and conclude that the dog correctly identifies cancer more than one fifth of the time.
 - There is enough statistical evidence to prove that the dog correctly identifies cancer more than one fifth of the time.
 - 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.
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?
- Observational study
 - Randomized experiment
 - Survey
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?

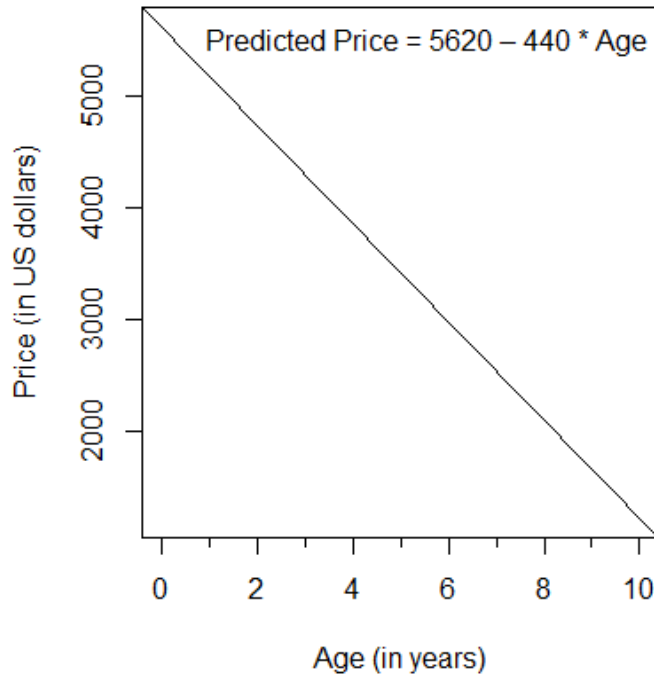
- Although 5,000 students live in dormitories on campus, only 500 were sent the survey.
- The survey was sent to only first-year students.
- Of the 500 students who were sent the survey, only 160 responded.
- All of the above present a problem for generalizing the results to all dormitory students at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

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



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

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?

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

Appendix L

Table L

Tetrachoric Correlation Residuals from the Single-factor Confirmatory Factor Analysis

Item	1	2	3	4	5	6	7	8
2	-0.001							
3	-0.015	0.015						
4	0.118	-0.007	0.024					
5	0.081	0.047	-0.011	0.038				
6	0.066	-0.046	0.022	0.071	-0.077*			
7	0.046	-0.014	0.098	0.127	-0.04	-0.047		
8	-0.038	-0.161	-0.029	0.024	0.024	0.08	0.028	
9	0.009	0.005	0.06	-0.103	0.004	0.072	0	-0.052
10	0.074	-0.009	-0.018	0.203	0.024	-0.034	0.015	-0.002
11	0.001	0.082	-0.009	-0.049	0.073	-0.055	-0.009	0.068
12	0.05	-0.01	0.06	-0.021	0.035	0.013	0.001	0.016
13	0.036	-0.05	0.014	0.105	-0.025	0.092	-0.041	0.037
14	-0.056	-0.051	0.086	0.067	0.066	-0.071	0.097	0.023
15	0.07	-0.048	-0.035	0.084	0.095	-0.007	0.014	0.067
16	0.083	-0.03	-0.019	0.128	-0.084	0.024	0.07	0.012
17	-0.005	0.025	0.006	-0.059	0.009	-0.03	0.033	0.112
18	-0.04	-0.014	-0.051	-0.007	-0.059	-0.044	-0.012	-0.032
19	-0.021	0	-0.045	-0.084	-0.009	-0.076	-0.079	-0.037
20	0.082	0.034	-0.03	-0.062	0.067	-0.005	0.036	0.051
21	0.001	-0.018	-0.043	0.073	-0.074	0.053	0.119	0.035
22	-0.066	-0.011	-0.066	-0.175	-0.08	0.05	0.025	-0.069
23	-0.005	-0.018	-0.009	-0.091	-0.077	0	-0.008	0.04
24	0.112	0.156	-0.116	0.064	-0.024	0.037	0.023	-0.095
25	-0.11	0.047	-0.037	-0.14	-0.086	0.003	-0.06	0.024
26	-0.039	0.012	0.076	-0.027	0.069	-0.048	-0.018	0.019
27	-0.018	-0.001	0.018	-0.083	-0.072	-0.033	0.054	-0.033
28	-0.074	-0.022	-0.04	-0.061	-0.051	0.076	-0.122	-0.019
29/30	-0.057	-0.037	0.013	0.02	-0.009	-0.026	-0.055	0.062
31	-0.073	-0.032	0.046	-0.112	-0.033	0.076	-0.006	-0.007
332	-0.001	0.004	-0.03	-0.036	0.114	0.02	-0.003	-0.108
33	-0.03	-0.073	0.017	-0.033	-0.113	0.01	-0.058	0.004
34	0.055	0.064	-0.054	0.121	-0.031	0.074	-0.041	-0.096
35	-0.098	0.035	0.008	-0.179	0.069	0.027	0.021	-0.025
36	-0.127	0.009	-0.034	-0.001	0.056	-0.049	-0.052	0.071
37	-0.02	-0.038	0.082	-0.01	0.088	-0.112	-0.041	0.001

(continued)

Note. * denotes a testlet

Table L (continued)

Item	9	10	11	12	13	14	15	16
10	0.074							
11	0.002	0.049						
12	-0.001	0.076	0.077					
13	0.068	-0.014	-0.118	0.008				
14	0.038	-0.033	0.055	0.019	-0.017			
15	-0.019	0.086	0.055	-0.058	0.011	0.055		
16	-0.052	-0.049	-0.035	0.041	0.029	0.107	0.05	
17	0.004	-0.015	0.068	-0.006	0.049	-0.026	-0.035	0.007
18	-0.089	0.066	-0.025	-0.049	0.05	0.02	-0.035	0.074
19	0.032	0.078	0.023	0.035	-0.012	-0.086	0.017	-0.067
20	0.032	-0.051	0.025	-0.022	-0.118	0.054	0.017	0.047
21	0.058	-0.128	-0.011	0.001	0.099	0	-0.116	0.098
22	-0.018	-0.103	-0.031	-0.033	-0.023	-0.011	-0.069	-0.017
23	-0.01	-0.09	-0.072	-0.021	-0.043	-0.058	-0.063	-0.082
24	0.018	-0.014	-0.037	-0.017	-0.064	0.021	-0.074	0.035
25	-0.096	-0.016	-0.071	0.023	0.029	-0.034	-0.105	-0.152
26	0.021	0.019	-0.053	-0.068	0.037	0.048	-0.036	-0.109
27	-0.049	-0.039	-0.125	0.044	-0.084	-0.001	0.098	0.113
28	-0.019	-0.033	0.055	-0.034	0.009	0.015	0.11	-0.068
29/30	0.031	0.016	0.083	0.017	0.03	-0.015	-0.006	-0.111
31	-0.056	-0.184	-0.039	-0.065	0.019	-0.078	-0.108	-0.028
332	-0.052	-0.039	-0.029	-0.002	0.034	-0.025	0.001	-0.043
33	0.037	0.047	0.099	-0.034	-0.047	-0.043	-0.083	0.018
34	-0.002	0.039	-0.059	-0.02	-0.046	-0.028	-0.025	0.075
35	0.012	-0.12	-0.039	-0.003	0.027	-0.018	0.029	0.066
36	0.003	0.038	0.045	-0.033	0.035	-0.113	0.036	-0.083
37	-0.02	0.099	0.022	0.042	-0.024	-0.057	0.039	-0.036

(continued)

Table L (continued)

Item	17	18	19	20	21	22	23	24
18	0.105							
19	0.064	-0.065*						
20	0.013	0.045	-0.016					
21	-0.083	0.026	-0.123	0.083				
22	-0.008	-0.018	0.017	0.026	0.079			
23	-0.041	0.067	-0.063	-0.047	0.13	0.018		
24	0.007	-0.059	-0.028	0.016	-0.041	0.031	-0.079*	
25	-0.077	-0.01	0.073	-0.077	0.033	0.172	0.165	-0.08
26	-0.023	-0.015	0.067	-0.021	-0.04	-0.019	0.026	-0.054
27	-0.068	0.094	-0.056	0.022	0.069	0.032	0.013	-0.038
28	-0.003	-0.126	0.043	-0.012	-0.14	-0.036	-0.012	0.006
29/30	0.016	-0.006	0.047	-0.053	-0.115	-0.009	0.039	-0.009
31	-0.019	0.03	-0.036	-0.012	-0.039	-0.003	0.089	0.126
332	-0.062	-0.018	0.053	0.034	0.022	0.089	0.085	-0.036
33	0.026	-0.031	0.078	-0.071	-0.1	0.056	0.044	0.01
34	-0.116	0.034	-0.026	-0.011	0.043	-0.021	0.044	0.098
35	-0.052	0.037	-0.086	0.075	0.194	0.056	0.013	0.04
36	0.053	-0.122	0.12	-0.105	-0.154	0.043	-0.039	-0.022
37	0.06	0.022	0.04	-0.016	-0.071	-0.091	-0.019	-0.031

Item	25	26	27	28	29/30	31	32	33
26	0.012*							
27	-0.035	0.019						
28	0.076	0.005	-0.037					
29/30	0.136	0.04	-0.041	-0.003				
31	0.191	-0.009	-0.067	0.023	0.069			
332	0.023	-0.102	0.157	0.111	-0.047	-0.038		
33	0.056	0.021	-0.085	0.059	0.002	0.053	-0.03	
34	0.06	0.025	-0.004	0.026	-0.055	0.022	-0.002	-0.034
35	-0.071	0.038	0.051	-0.02	-0.085	0.052	-0.003	-0.008
36	-0.02	0.065	-0.169	0.108	0.142	-0.148	-0.153	0.063
37	-0.127	-0.002	0.021	0.04	0.024	-0.074	0.006	0.052

Item	34	35	36
35	0.021		
36	-0.002	-0.174	
37	-0.132	-0.077	0.187

Note. * denotes a testlet