

Developing and Validating an Instrument to Measure College Students' Inferential  
Reasoning in Statistics: An Argument-Based Approach to Validation

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Jiyoon Park

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Robert delMas, Adviser

Joan Garfield, Co-adviser

June 2012

© Jiyeon Park 2012

## Acknowledgements

*“Your word is a lamp to my feet and a light for my path.” (Psalms 119: 105)*

The journey towards a PhD would not have been possible without many people. I would like to express my deep thanks to esteemed co-promoter Dr. Bob delMas. As my academic advisor, he has directed my development as a scholar and researcher. I am truly grateful for his enthusiasm and support in every aspect over the years. I look forward to a new dynamic in our relationship as I pursue my own academic career.

I would like to thank Dr. Joan Garfield for the trust, the insightful discussion, offering valuable advice, for our support during the whole period of the study. I especially thank for your patience and guidance.

I would like to thank Dr. Michael Rodriguez for his excellent advises on designing the study, methods of data analysis, and integrating the results. I appreciate your clear guidance that made my dissertation completed successfully.

I have also benefited from a rich teaching and research environment during my graduate years in QME program at University of Minnesota. I thank to the students and faculty in statistics education group. I am proud of the tremendous research accomplishments our group has made in statistics education research. In particular, I would like to acknowledge great

To my family, my Mom, Mrs. Young-Ui Lee, my Dad, Mr. Jae-Hou Park, my brother, Mr. Jong-Hyun Park, and my sister-in-law, Mrs. Hyun-Jung Lee, I warmly and deeply thank for your loving-support. I am fully blessed by your influence in my life.

I want to express my special gratitude and deepest appreciation to my remaining family and friends too numerous to name for love, support, and prayer. All of you helped

me see the lights of Jesus and prayed for me as I am in the dark. I truly and deeply appreciate you all.

## Abstract

The purpose of this study was to develop and validate an assessment to measure college students' inferential reasoning in statistics. This proposed assessment aims to help statistics educators guide and monitor students' developing ideas of statistical inference.

Within the two-stage cycle, the formative and summative stages, this study first built arguments for the use of assessment and score interpretations, and verified inferences made from those arguments. The five claims were used to examine the plausibility of the validity arguments: 1) The test measures students' level of statistical inferential reasoning in two aspects—informal statistical inference and formal statistical inference; 2) The test measures statistical inferential reasoning in the representative test domains; 3) The test produces scores with sufficient precision to be meaningfully reported; 4) The test is functional for the purposes of formative assessment; and 5) The test provides information about students' level of statistical inferential reasoning in the realms of informal and formal statistical inference.

Using a mixed-methods study design, different types of validity evidence were gathered and investigated. Three content experts provided their evaluation of the test blueprint and assessment, based on their qualitative reviews. For the revised assessment resulting from the experts' feedback, cognitive interviews were conducted with nine college students using think-aloud protocols, whereby the students verbalized their reasoning as they reached an answer. A pilot-test administered in a classroom provided preliminary information of the psychometric properties of the assessment. The final version of the assessment was administered to 2,056 students in 39 higher education institutions across the United States. For the data obtained from this large-scale

assessment, a unidimensional model in confirmatory factor analysis and the Graded Response Model in item response theory were employed to examine the arguments regarding the internal structure and item properties. The results suggest that the AIRS is unidimensional with appropriate levels of item difficulty and information. The pedagogical implications for the use of the AIRS test are discussed with regard to the areas where students showed difficulties in the domain of statistical inference.

## Table of Contents

Acknowledgements	i
Abstract	iii
List of Tables	xi
List of Figures	xiii
Chapter 1 Introduction	1
Statistical Inference	1
Difficulties Understanding Statistical Inference	1
Informal Statistical Inference (ISI) and Formal Statistical Inference (FSI)	2
New Instructional Approaches to Develop Students' Understanding of Statistical Inference	4
Need for New Assessments	5
Overview of the Study	6
Overview of the Chapters	7
Chapter 2 Review of the Literature	10
What is Statistical Inference?	11
Definition and Importance of Statistical Inference	11
Paradigms of Statistical Inference	12
Fisherian versus Neyman-Pearson	14
Controversies about Null Hypothesis Statistical Testing (NHST)	15
Research Studies on Inferential Reasoning in Statistics	19

Studies on Foundations of Statistical Inference	20
Studies about Formal Statistical Inference	28
What is Informal Inferential Reasoning (IIR)?	34
Inferential Reasoning in Statistics (IRS)	35
Definition and Components of IIR	36
Role of IIR in Reasoning about Statistical Inference	39
Studies on Informal Statistical Inference	40
Content Domains of Statistical Inference	44
Need for an Instrument to Assess Inferential Reasoning in Statistics	46
New Instructional Approaches to Develop Students’	
Understanding of Statistical Inference	46
Existing Assessments	47
Why These Assessments Do Not Meet the Current Need	48
Summary of the Literature Reviewed	49
Formulation of the Problem Statement	51
Chapter 3 Methods	54
Validity and Validation	54
Validity	54
A Validation Method: An Argument-based Approach to	
Validation	58
Framework of the Study	60
Formative Stage: Formulating the Interpretive Argument and	
Assessment Development	61



Developing the Interpretive Argument	61
Developing a Test Blueprint from the Literature Review (Theoretical Evidence 1: TE1)	63
Expert Review of the Preliminary Test Blueprint (Theoretical Evidence 2: TE2)	64
Test Specifications (Theoretical Evidence 3: TE3)	66
Developing an Item Pool (Theoretical Evidence 4: TE4)	69
Expert Review for the Preliminary Assessment (Theoretical Evidence 5: TE5)	69
Summative Stage: Validating the Interpretive Argument	70
First cognitive Interview Using Think-alouds (Empirical Evidence 1-1: EE1-1)	70
A Pilot-test (Empirical Evidence 2: EE2)	75
Second Cognitive Interview Using Think-alouds (Empirical Evidence 1-2: EE 1-2)	75
Field-testing (Empirical Evidence 3: EE3)	78
Chapter 4 Results	91
Analysis Results for the Data Obtained in the Formative Stage	91
Results from the Literature Review to Create the Test Blueprint: Theoretical Evidence 1 (TE1)	91
Expert Review of the Preliminary Test Blueprint: Theoretical Evidence 2 (EE2)	95
Test Specifications: Theoretical Evidence 3 (TE3)	101

Examining Existing Instruments and Literature for Developing	
Preliminary Test: Theoretical Evidence 4 (TE4)	102
Expert Review for the Assessment Items: Theoretical Evidence	
5 (TE 5)	105
Analysis of Results in the Summative Stage	110
First Cognitive Interview: Empirical Evidence 1 (EE1)	110
Results from Pilot Testing: Empirical Evidence 2 (EE2)	120
Second Cognitive Interview: Empirical Evidence 3 (EE3)	124
Results from Field-testing: Empirical Evidence 4 (EE4)	128
Synthesis of the Results	148
Evaluation of Scoring Inference	148
Chapter 5 Summary and Discussion	159
Summary of the Study	159
Discussion of the Claims	162
Is IRS Unidimensional or Multi-dimensional?	163
How Useful is this Instrument?	164
Limitations	165
Teaching Implications	166
Implications for Future Research	169
Conclusion	169
References	171
Appendix A. Studies on Statistical Inference	199
Appendix B. Preliminary Test Blueprint	207

Appendix C. Expert Review Forms of Test Blueprint	212
Consent Form: Expert Review	212
The Invitation Letter and Test Blueprint Evaluation Form	214
Test Blueprint Evaluation Form	215
Appendix D. Final Version Test Blueprint	216
Appendix E. Expert Review Forms of Preliminary Assessment	220
Item evaluation form (general)	220
Item Evaluation Form (specific)	221
Appendix F. Student Cognitive Interview Invitation	222
Student Invitation Letter: Cognitive Interview	222
Consent Form: Student Cognitive Interview	223
Appendix G. Online Assessment Consent Form and Test Instruction	225
Appendix H. Expert Review on Test Blueprint	227
Appendix I. Versions of Assessment	234
Preliminary Version	234
AIRS-1 (Changes were made from expert reviews)	246
AIRS-2 (Changes were made from 1st cognitive interview)	258
AIRS-3: Final version (Changes were made from pilot testing)	270
AIRS Online Consent Form	270
Start AIRS	270
Quiz Score	281
Appendix J. Expert Review on Preliminary Assessment	282
Appendix K. Reasoning Statement and Expert’s Enacted Reasoning	286

Appendix L. Reliability Analysis from Pilot Testing	292
Appendix M. LD Indexes of AIRS Items	294
Appendix N. Development of a Preliminary Version: Item Changes Made from Existing Instruments	298

## List of Tables

Table 1. <i>Think-aloud Coding Framework</i>	75
Table 2. <i>Changes in the AIRS Versions</i>	78
Table 3. <i>Types of Institution in a Large-Scale Assessment</i>	80
Table 4. <i>Mathematics Pre-requisites for the Statistics Course</i>	80
Table 5. <i>Structure of the Testlet-Based Test</i>	83
Table 6. <i>Contingency Tables of Observed- and Expected Frequency</i>	85
Table 7. <i>Summary of Data Collection Phases</i>	89
Table 8. <i>Examples of the Preliminary Blueprint</i>	93
Table 9. <i>Results of Expert Review on Test Blueprint</i>	96
Table 10. <i>Changes to Test Blueprint Implemented from Expert Reviews</i>	99
Table 11. <i>Resources of Items in a Preliminary Version</i>	104
Table 12. <i>Items rated "Strongly Disagree" or "Disagree" by at least One Reviewer</i>	105
Table 13. <i>Changes made for the Items Rated "Strongly Disagreed" or Disagreed"</i>	107
Table 14. <i>Excerpts of Expert's In-depth Cognitive Interview: Selected Notes</i>	112
Table 15. <i>Excerpts of Students' 1st Cognitive Interview: Selected Notes</i>	117
Table 16. <i>Item Difficulties (Proportion Correct) of AIRS Items</i>	121
Table 17. <i>Changes Made in AIRS-3 from Pilot-testing</i>	124
Table 18. <i>Coding Categories Made for Cognitive Interviews</i>	126
Table 19. <i>Results of Coding Cognitive Interviews</i>	128
Table 20. <i>Coefficient-alpha Reliabilities</i>	129

Table 21. <i>Mean LD Indices of Each Item</i>	132
Table 22. <i>Factor Loadings</i>	135
Table 23. <i>Fit Indices for Factor Models</i>	138
Table 24. <i>Results of Fitting a GRM Model</i>	141
Table 25. <i>Item Difficulties as Proportion-correct</i>	167
Table A-1 <i>Studies on Foundations of Statistical Inference, Formal Statistical Inference, and Informal Statistical Inference</i>	199
Table B-1. <i>Test Blueprint to Assess Informal Statistical Inference</i>	207
Table B-2. <i>Test Blueprint to Assess Formal Statistical Inference</i>	209
Table D-1. <i>Test Blueprint to Assess Informal Inference</i>	216
Table D-2. <i>Test Blueprint to Assess Formal Inference</i>	218
Table H-1. <i>Summary of Expert Comments</i>	227
Table H-2. <i>Detailed Comments</i>	231
Table J-1. <i>Comments of Reviewers</i>	282
Table K-1. <i>Reasoning statement (intended reasoning) in AIRS-1</i>	286
Table K-2. <i>A Script from an Expert's Think-aloud</i>	289

## List of Figures

<i>Figure 1.</i> Conjectured relationships between the terms related to statistical inference.	36
<i>Figure 2.</i> Kane, Crooks, & Cohen (1999).	59
<i>Figure 3.</i> Distribution of total scores in pilot-test.	120
<i>Figure 4.</i> Q-Q plot of correct-total scores in pilot-test.	122
<i>Figure 5.</i> Item characteristic curves of 19 testlet-based items.	145
<i>Figure 6.</i> Item information curves of 19 testlet-based items.	146
<i>Figure 7.</i> Test information function and standard error of measurement.	147
<i>Figure 9.</i> Q-Q plot of correct-total scores.	150
<i>Figure 10.</i> Distribution of IRT scores.	151
<i>Figure 11.</i> Q-Q plot of IRT scores.	151
<i>Figure 12.</i> Scatter plot of correct-total scores (34 items) versus IRT scores.	152

## Chapter 1

### Introduction

#### **Statistical Inference**

In David Moore's textbook (2007), statistical inference is described as "moving beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are therefore uncertain" (p. xxviii). Garfield and Ben-Zvi (2008) grouped the topics of statistical inference into two categories, parameter estimation and hypothesis testing.

The ability to draw inferences from data is a part of everyday life as people are confronted with situations where they need to critically review data-based claims (Garfield & Ben-Zvi, 2008). Understanding of statistical inference is important in scientific research since the concepts and processes in statistical inference are used in all empirical studies (Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007).

In introductory statistics courses, students learn hypothesis tests and confidence intervals as main methods of making conclusions for quantitative data. A learning goal of the college-level Guidelines for Assessment and Instruction in Statistics (GAISE; ASA, 2005) is that students develop and evaluate inferences and predictions that are based on data. The GAISE report recommends that students should understand the basic idea of statistical inference, and emphasize the concept of a sampling distribution and how it applies to making statistical inferences.

#### **Difficulties Understanding Statistical Inference**

There seems to be an agreement about the importance of statistical inference (e.g., Aberson, Berger, Healy, & Romero, 2003; Garfield & Ben-Zvi, 2008). However, many



misunderstandings have been reported that people are confused about the concepts and processes in statistical inference (Falk & Greenbaum, 1997; Haller & Kraus, 2002; Wilkerson & Olson, 1997; Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). For example, Tversky and Kahneman (1971) showed that people believe that any sample must be similar to the population, regardless of its sample size. After this work, Kahneman and Tversky established a cognitive basis for common human errors people show in statistical inference.

More recently, there have been studies about people's difficulty understanding hypothesis testing. Specifically, research has revealed that students have difficulty understanding—the definition of the hypotheses (Vallecillos & Batanero, 1997), the definition of significance level and the  $p$ -value (Falk, 1986), and the logic of hypothesis testing (Vallecillos, 1999). Regarding students' difficulties understanding formal statistical inference, research studies have been conducted on why people show those misunderstandings. Several studies have been conducted about difficulties students have understanding concepts in sampling distribution (e.g., Chance, delMas, & Garfield, 2004; Saldanha & Thompson, 2002), which is a foundational concept to understand statistical inference. Some studies have shown that students do not differentiate between the distribution of a sample and the sampling distribution of a statistics (e.g., Lipson, 2003). There are also studies that have revealed students' difficulty understanding the concepts involved in the Central Limit Theorem (e.g., Batanero, Tauber, & Sanchez, 2004).

### **Informal Statistical Inference (ISI) and Formal Statistical Inference (FSI)**

In the past few years, statistical educators have looked for new ways to help students build an understanding of statistical inference, in light of current research and

new developments in the practice of statistics. As a way to support a coherent understanding of the concepts and processes in statistical inference, Wild, Pfannkuch, Regan, and Horton (2011) suggest a learning pathway that introduces some of the “big ideas” behind inference before teaching *formal statistical inference*. Garfield and Ben-Zvi (2008) address that ideas of inference should be introduced informally at the beginning of the course, such as having students become familiar with seeing where a sample corresponds to a distribution of sample statistics, based on a theory or hypothesis. They further argue that this may help students be less confused by the formal ideas, procedures, and language when they finally reach the formal study of this topic.

The big ideas of inference that can be taught before formal inference, suggest two content areas in statistical inference—informal statistical inference (ISI) and formal statistical inference (FSI). In this paper, these terms are used to specifically refer to the *content areas of* statistical inference. The topics of ISI include: the concept of uncertainty; properties of aggregate data; recognizing sampling variability; the concept of unusualness; (informal) generalization from a sample to a population; (informal) comparison between two samples. The concepts involved in formal statistical testing (e.g.,  $p$ -value, statistical significance, hypothesis tests, confidence intervals) are categorized as FSI. In addition, the topics of foundations of formal statistical inference (e.g., sample representativeness, sample variability, sampling distribution) are also included in this category given that they are foundational to understanding formal statistical inference (e.g., Chance et al., 2004).

Although there has been increased attention given to informal ideas in statistics, it is only recently that researchers and educators attempted to characterize the distinctive

features of Informal Inferential Reasoning (IIR). For example, recent forums of the International Research Collaboration on Statistical Reasoning, Thinking and Literacy (SRTL-5, 6, and 7 in 2007, 2009, 2011, respectively), have gathered statistics education researchers to discuss and share their research on IIR. Particularly at SRTL-5 (2007), statistics education researchers put their efforts to characterize the nature of informal reasoning through exploratory studies. Published articles (Ben-Zvi, 2006; Pfannkuch, 2006a; Pfannkuch, 2006b; Pratt, Johnston-Wilder, Ainley, & Mason, 2008; Zieffler, Garfield, delMas, & Reading, 2008) share a common understanding about IIR represented with three principles: (1) generalizations that go beyond describing the given data; (2) the use of data as evidence for those generalizations; and (3) conclusions that express a degree of uncertainty, quantified or not, accounting for the variability or uncertainty.

### **New Instructional Approaches to Develop Students' Understanding of Statistical Inference**

Accompanied with ongoing calls for reform in introductory statistics courses, different teaching methods for developing students' inferential ideas have been proposed. For example, Cobb (2007) and Kaplan (2009) have suggested major changes in how statistical inference is taught in the introductory college course. Cobb (2007) challenges statistics educators to purposefully reconsider and the content of introductory statistics courses. Cobb (2007) argues, flexible and accessible technological tools now allow the logic of inference to be put at the core of introductory course instead of the normal distribution. Statistical inference can now be taught using a randomization approach (e.g., permutation tests) instead of asymptotic sampling distributions. Cobb (2007) suggests

that using permutation tests to learn statistical inference provides students with both a conceptually easier instruction to statistical inference and a modern, computational data analysis technique currently lacking in the first course in statistics. Similar to Cobb, Kaplan (2009) also suggested that resampling or permutation methods reflect more easily generalizable ideas and, for many students, they are more accessible conceptually.

Inspired by Cobb (2007) and Kaplan (2009), recent NSP-sponsored projects have developed new curriculum to introduce students to ideas of statistical inference using randomization methods (e.g., The CSI project, headed by Rossman, Chance, Cobb, & Holcomb (<http://statweb.calpoly.edu/csi>); The CATALST course, developed by Garfield, delMas and Zieffler (<http://www.tc.umn.edu/~catalst>); The INCIST project, headed by West, 2011).

### **Need for New Assessments**

Now that there is increasing attention to randomization-simulation based curricula to help students better understand statistical inference, there is a need to investigate the impact of these curriculum on student learning and understanding of statistical inference. In addition, despite increased interest in informal inferential reasoning and efforts to characterize IIR, there are no assessments of IIR or studies on how IIR relates to reasoning about formal statistical inference.

There are existing instruments used in statistics education research and evaluation to measure students' reasoning in statistics (e.g., The Statistical Reasoning Assessment (SRA), Garfield, 1998; The Statistics Concepts Inventory (SCI), Reed-Rhoads, Murphy, & Terry, 2006; and the Assessment Resource Tools for Improving Statistical Thinking (ARTIST), Garfield, delMas, & Chance, 2002). Although these instruments assess

important outcomes (e.g., assessing students reasoning, thinking, and conceptual understanding), the topics assessed in these instruments do not cover the full domain of reasoning about statistical inference. Thus, these existing instruments do not align with the current needs of an assessment: measuring informal inference in association with reasoning about formal inference; and assessing inferential reasoning of students taught with randomization-simulation methods.

Moreover, the existing instruments have not been developed or validated using modern psychometric measurement models (e.g., item response theory) that provide ample information about properties of items (e.g., item difficulty, item discrimination, item information). Therefore, there is also a need for a new instrument that is developed and validated using modern psychometric theory so that the results from the assessment provide reliable and valid interpretations.

### **Overview of the Study**

In response to the need for a new research instrument, this study was designed to develop a reliable and valid measure to assess college students' inferential reasoning in statistics (IRS). In this study, IRS is defined as the way people draw conclusions from data at hand to a broader context using the concepts and ideas of statistical inference.

This instrument will eventually allow several questions to be addressed in future studies: How do students use informal ideas to understand formal concepts in statistical inference? What kinds of informal ideas do students have before learning formal ideas? How are these two types of inference related each other in students' reasoning process?

This study attempts to build and support arguments for the use of the assessment of evaluating informal and formal statistical reasoning of students in introductory

statistics courses. An argument-based approach to validity (Kane, 1992, 2001, 2006a, 2006b; Kane, Crooks, & Cohen, 1999) was employed as a way to justify its score-based interpretations and uses as an overarching logical framework. This approach guided the development of the assessment and validation of the interpretive arguments in an iterative process between test development and validation.

This study was structured with two stages, following Kane's framework: a formative and summative stage. In the formative stage, interpretive arguments were specified based on claims regarding the proposed test score interpretations and uses. A test blueprint and assessment were developed at this stage. A review of the literature was used to develop the preliminary versions of test blueprint and assessment. Expert reviews were used to revise the preliminary version of the test blueprint and the assessment. Those sources also provided theory-based evidence to support the interpretive arguments.

At the summative stage, different types of empirical evidence were collected and examined. As evidence sources, cognitive interviews with an expert and students, pilot testing, and field-testing for large-scale assessment were gathered. Interpretative arguments were investigated in terms of their plausibility by examining the extent to which each kind of evidence supports the claims underlying the arguments.

### **Overview of the Chapters**

This thesis includes five chapters. The current chapter, Chapter 1, provides background on current perspectives in statistics education, and an overview of the research problem, focusing on the research on difficulties understanding statistical inference and drawing on the need for an instrument to measure IRS. Chapter 2 reviews the literature related to IRS. This chapter provides the theoretical perspectives of major

inferential statistical concepts and tools (e.g., *P*-value, hypothesis tests), as well as controversies on the use of those tools. Relevant previous research studies on the topic of statistical inferential reasoning are examined. Research studies are reviewed on foundations of statistical inference and formal statistical inference. This chapter also reviews studies about IIR in terms of definition and components of IIR. Key findings from the major studies on the topic of IIR are also reviewed. Existing instruments to assess students' reasoning in statistics are examined to inform the need for a new instrument to measure students' inferential reasoning in statistics.

Chapter 3 describes the methodology used in this study. A description of validity and validation methods (an argument-based approach to validation by Kane) is provided with a framework of the study. Claims regarding the proposed assessment are then provided specifying what to measure and how to use the test results. This set of claims plays an enabling role supporting an interpretative argument as different types of evidence are investigated. Different kinds of evidence to support validity arguments are described. This chapter describes the formative stage and summative stage of instrument development and validation, and in each stage, different kinds of evidence sources are explained with information on study participation, methods of data collection, and analysis methods.

Chapter 4 reports the outcomes of the assessment development and validation. With the same structure as Chapter 3, the evidence sources collected in each stage are examined to evaluate the plausibility of the claims. After all the evidence sources are investigated, it synthesizes the research arguments, considering all aspects of the analysis results. Underlying inferences about test uses and score interpretations are evaluated by

judging the claims laid out in the formative stage. Finally, Chapter 5 provides a summary of the research findings and discusses the research and teaching implications. This chapter also includes a discussion of future research.



## Chapter 2

### Review of the Literature

This chapter describes the literature that is relevant to statistical inference. The review begins with definitions of inference and statistical inference. Historical paradigms of statistical inference are summarized with respect to how probability has been interpreted. Issues with the application and interpretation of statistical testing follow, including a discussion of two different approaches to hypothesis testing. Debates regarding null hypothesis statistical testing are then followed.

Next, research studies about statistical inference are presented with two subsections: foundations of statistical inference and formal statistical inference. Studies on foundations of statistical inference are centered around literature on reasoning about sampling distribution considering that the concept of sampling distribution represents an important building block to a coherent understanding of statistical inference (Chance et al., 2004; Noll, 2011). Reviews on literature about the topics of formal statistical inference, such as hypothesis testing, are then described. Methodologies used, major findings, the inferences made from the results, and the implications are examined. A literature review about informal inferential reasoning is then presented in terms of its background, definitions and characteristics. Recent studies conducted on informal inferential reasoning are reviewed.

From the research studies reviewed, a domain of statistical inference is categorized into two content categories—formal statistical inference (FSI) and informal statistical inference (ISI). Research questions are posed in order to inform what research has not yet answered.

## What is Statistical Inference?

### Definition and Importance of Statistical Inference

Moore (2007) states that statistical inference “moves beyond the data at hand to draw conclusions about some wider universe, *taking into account* that variation is everywhere and the conclusions are uncertain” (p.172). Moore’s perspective gives a general idea about statistical inference; making a conclusion about an uncertain, broader context from data.

The ideas of statistical inference are used in all empirical sciences (Sotos et al., 2007). Saldanha and Thompson (2007) note that ideas of sampling and statistical inference are important to understand “the degree to which data-based claims are warranted” and to understand that “conflicting claims are not necessarily a sign of confusion or duplicity” (p. 271). In the field of statistics education, it is clear that statistical inference is a necessary skill in everyday citizenship. Garfield and Ben-Zvi (2008) note that drawing inferences from data is a part of everyday life, and critically reviewing the results of statistical inferences from research is an important capability for all adults.

The 2000 Curriculum standards for grades 6-12 mathematics state that all students should develop and evaluate inferences that are based on data. With regard to teaching statistical inference, the NCTM standards include recommendations for grades 9 to 12—students should:

- Use simulations to explore the variability of sample statistics from a known population and to construct a sampling distribution.

- Understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference. (p. 324, NCTM, 2000)

The Guidelines for Assessment and Instruction in Statistics (GAISE) report (ASA, 2005), a document to provide a conceptual framework for K-12 statistics education, recommends that students develop and evaluate inferences and predictions that are based on data. In the GAISE report at the college level, statistical inference is considered to be more important. Understanding the ideas of statistical inference is regarded as the most important learning goal in introductory statistics course. The GAISE report emphasizes understanding the concept of a sampling distribution and how it applies to making statistical inferences, based on samples of data (including the idea of standard error); the concept of statistical significance, including significance levels and *P*-values; and the concept of confidence intervals, including the interpretation of confidence levels and the margin of error. Therefore, it is evident that reasoning about statistical inference is necessary skill in everyday life, and the concepts and ideas of statistical inference have been emphasized in school curricula.

### **Paradigms of Statistical Inference**

The application of formal statistical methods stems from different historical paradigms as well as psychologists' reconciliation between two different approaches to use of the methods (Halpin & Stam, 2006). Understanding of these historical backgrounds allows statistics educators to help students better learn and apply statistical concepts and methods with comprehensive view on the ideas.

There are three interpretations of probability that affect statistical inference: the classical, frequentist, and Bayesian approaches. In the classical approach, the probability is understood as the ratio of the number of alternatives favorable to that event to the total number of equally-likely alternatives (Konold, 1991). This approach has been criticized in that this interpretation is limited to trials with objects such as coins, dice and spinners, which are composed of equally-likely alternatives.

The frequentist approach emerged as a way to address the paradoxes of the classical approach. In this view, a probability represents a long-run frequency by considering repeated sampling of datasets similar to the one at hand (Cox, 2005). Kyburg (1974) notes that the most desirable probability is one that tells us how to anticipate the future perfectly. In this sense, the most attainable and simplest rule is to ignore the arithmetic and act “as we feel like acting” (p.23). Kyburg addresses that the need for statistical inference comes from a situation where we are uncertain about how to behave under certain circumstances (Kyburg, 1974). Along the same lines, Lehman’s (1991) view on probability begins with uncertainty, and he addresses that data from observations provide guidance as to the best decision for the uncertain situation.

The last paradigm in understanding formal statistical inference is the Bayesian approach. From this view, probability is “a degree of belief held by a person about some hypothesis, event, or uncertain quantity” (Phillips, 1973, as cited in Cox, 2005). Instead of using probability as representing a long-run frequency, the Bayesian approach attempts to attach a probability distribution to the unknown probability distribution. In other words, Bayesian inference uses available posterior beliefs as the basis for making statistical propositions.

## **Fisherian versus Neyman-Pearson**

The frequentist approach dominated the uses and application of statistics in scientific research between 1940 and 1960, and many statistical methods were developed in this period. Halpin and Stam (2006) further discern this period by considering extant disagreements about the application and interpretation of statistical testing, represented by two opposing theories R. A. Fisher and J. Neyman-E. S. Pearson propounded (Halpin & Stam, 2006). The debates between Fisher and Neyman-Pearson (N-P) come out from their different perspectives on hypothesis significance testing: a Fisherian test involves only one hypothesized model, whereas an N-P test involves two hypotheses, a null hypothesis and an alternative hypothesis. In Fisherian tests, the distribution of the data must be known, and this distribution is used both to determine the test and to evaluate the outcome of the test. On the other hand, in the N-P perspective, the researcher chooses a null hypothesis and tests the null against the alternative hypothesis (Christensen, 2005, p. 121).

Batanero (2000) notes that the test forms and results from the tests are nearly the same, but the underlying philosophy and the interpretation of the results are profoundly different. She states that the philosophical basis of a Fisherian test is “proof by contradiction” since Fisherians confront a null hypothesis with observations, and a *P*-value indicates the strength of the evidence against the hypothesis. For this reason, the Fisherian approach is referred to as a “test of significance” rather than a “test of hypothesis.”

In the procedure of significance testing, the *P*-value gives a measure of the extent to which the data do not contradict the model (Hubbard & Bayarri, 2003). Fisherians

interpret a  $P$ -value as the probability of seeing weird data rather than the probability of rejecting the null. On the contrary, in the N-P approach, a statistical test is a rule of “inductive behavior”—a criterion for decision-making that allows us to accept or reject a hypothesis (Christensen, 2005). In this case, the problem of statistical hypothesis testing occurs when we need to make a choice between two competing courses of action (Batanero, 2000).

There have been extensive debates between these two approaches. The critics of Fisherians argue that, if the model is not rejected, the best interpretation for the result from significance testing is that “the data are consistent with the model” (Christensen, 2005, p.122). In other words, since not rejecting the model certainly does not prove that it is correct, the interpretation of nonsignificant outcomes from significance testing is ambiguous in Fisherian approach (Halpin & Stam, 2006; Hubbard & Bayarri, 2003). The N-P approach has been criticized mostly because of its misuse and misinterpretation of results in practice. Critics of the N-P approach argue that it focuses on a small  $\alpha$ -level; thus, it often leads to bad decisions between the two alternative hypotheses.

### **Controversies about Null Hypothesis Statistical Testing (NHST)**

Null Hypothesis Statistical Testing (NHST) has arguably been the most widely used method of data analysis for the past 70 years (Nickerson, 2000). One great appeal of NHST is that it provides the use of “a straightforward, relatively simple method of extracting information from noisy data” (Wainer & Robinson, 2003, p. 28). It is also considered to be “an objective, scientific procedure of advancing knowledge” (Kirk, 2001, p. 214). Although NHST has served an important purpose in the advancement of scientific study inquiry, there have been debates regarding the use of NHST.

Several statisticians, as well as educators, criticize NHST partly because of its nature (e.g., Cohen, 1994; Falk & Greenbaum, 1995; McDonald, 1997; Rosnow & Rosenthal, 1989) and also because of its misuse and misinterpretation (e.g., Cohen, 1994; Falk, 1986; Falk & Greenbaum, 1995; Gigerenzer, 1993; Sedlemeier & Gigerenzer, 1989; Thompson, 1989, 1996). Cohen (1994) provides a review of the problems of NHST, as well as its misinterpretation. He points out the logical flaw of “deductive syllogistic reasoning” embedded in NHST. The basic structure of the NHST is—*If the  $H_0$  is correct, then these data are highly unlikely. These data have occurred. Therefore, the  $H_0$  is highly unlikely* ( $H_0$  is probably not true, and therefore, formally invalid). A misapplication of this “deductive syllogistic reasoning” is also pointed out by Falk and Greenbaum (1995). They call the logic behind NHST an “illusion of probabilistic proof by contradiction”. Cohen further argues that NHST does not tell us “what we want to know,” but rather tells us, “Given that  $H_0$  is true, what is the probability of these data?” (p. 997). Kirk (2001) also criticizes NHST in that it does not tell us how large the effect is, or whether the effect is important or useful.

In addition to these flaws in the nature in NHST, several researchers have considered the misuse and misinterpretation of NHST. The following are misunderstandings regarding the interpretation of NHST that have been most often addressed in a literature review of NHST uses.

- Misbelief that failing to reject the null hypothesis is equivalent to demonstrating it to be true (Batanero, 2000; Nickerson, 2000).

- Misbelief that the  $P$ -value is the probability that the null hypothesis is true, and that  $(1-p)$  is the probability that the alternative hypothesis is true (Carver, 1978; Falk & Greenbaum; 1995; Nickerson, 2000).
- Misbelief that a small  $P$ -value means a treatment effect of large magnitude (Cohen, 1994; Rosenthal, 1993).
- Misbelief that a small  $P$ -value is evidence that the results are replicable (“replicability fantasy”; Carver, 1978; Falk & Greenbaum; 1995; Gigerenzer, 1993; Greenwald, 1975; Rosnow & Resenthal; 1989; Thompson, 1996).
- Confusion between “significant” and “statistically significant” (Meehl, 1997; Thompson, 1996; Schafer, 1993).

Why are these confusions about NHST so pervasive? The most plausible explanation of this comes from two incompatible origins of statistical testing—Fisher and Neyman-Pearson—described in the previous section. Batanero (2000) argues that the current practice of statistical tests contains elements of decision procedures from N-P but elements of inferential procedures from Fisher. She notes that these two approaches “[are applied] at different stages of the process” (p. 87), although they are not comparable (Christensen, 2005; Gigerenzer, 1989). The significance of this hybridization of two different views has also been described as “a failure to understand the foundations of statistical inference” by Hubbard and Bayarri (2003, p.171). Similarly, Gigerenzer, Swijtink, Porter, Daston, Beatty, and Kruger (1989) maintain that the dispute between the two views has been hidden in applications of statistical inference in psychology and other experimental sciences, in which it has been assumed that there is only one statistical



solution to inference. Christensen (2005) provides a further argument on the incompatibility of the Fisherian and N-P approaches:

Many of them [the N-P testers] tend to adopt the philosophy of Fisherian testing (involving *P*-values, using small alpha levels, and never accepting a null hypothesis) while still basing their procedure on an alternative hypothesis....The motivation for using small alpha levels seems to be based entirely on the philosophical idea of proof by contradiction. Using a large alpha level would eliminate the suggestion that the data are unusual and thus tend to contradict  $H_0$ . However, N-P testing cannot appeal to the idea of proof by contradiction. (p. 123)

With regard to incomparable ideas between the Fisherian and N-P approaches, Wainer and Robinson (2003) provide Fisher's original idea of statistical testing:

When *p* is small, [Fisher] declared that an effect has been demonstrated. When it is large, he concluded that, if there is an effect, it is too small to be detected with an experiment this size. When it lies between these extremes, he discussed how to design the next experiment to estimate the effect size. (p. 23)

This indicates that the current practice of usage and interpretation of NHST is far from Fisher's original idea, which considers a *P*-value as the strength of evidence against the hypothesis, as opposed to a decisive tool for making a decision between dichotomous hypotheses.

In order to improve the current practices of NHST, some suggestions have been presented. First, NHST can be a valued tool when accompanied by effect sizes that

provide information regarding the trustworthiness of estimates of the effect size (e.g., Cohen, 1994; Wainer & Robinson, 2003). Kirk (2001) notes that the focus of research should be on “what the data tell us about the phenomenon under investigation” rather than on rejecting a null hypothesis and obtaining a small *P*-value (p. 213). Wainer and Robinson (2003) also note that NHST is most often useful as an adjunct to other results (e.g., effect sizes) rather than as a stand-alone result. Similarly, Schmidt (1996) argues that confidence intervals offer a solution for many problems associated with the use of NHST. In addition to combining information on location and precision, confidence intervals are considered as a tool to convey information on effect size (Schmidt, 1996; Cohen, 1994), as well as to reduce binary thinking (Hoekstra, Kiers, & Johnson, 2010). Furthermore, a confidence interval is considered to be easier to interpret, insofar as it is a visual representation of effect size and a measure of uncertainty (Schmidt & Hunter, 1997); thus, both can be seen at a single glance (Hoekstra et al., 2010).

Although there have been different historical paradigms of statistical inference (classical, frequentist, and Bayesian) and debates on the use of hypothesis testing, this study focuses on the methods currently being taught in most of the introductory statistics courses—Neyman-Pearson approach in frequentist perspective.

### **Research Studies on Inferential Reasoning in Statistics**

A review of research literature is structured into two subsections—studies about foundations of statistical inference and studies about formal statistical inference. This structure is reflected in the content and order of topics shown in most textbooks of introductory statistics courses (e.g., Moore & McCabe, 2006). In these textbooks, samples and sampling distributions, and the central limit theorem are explained as

foundations to inferential statistics. Students then learn how to perform formal statistical testing such as hypothesis tests.

Given that understanding sampling distributions is regarded as foundational to an understanding of formal statistical inference, review of the literature on the foundations of statistical inference is focused on studies about understanding sampling distributions. The second category of literature review includes studies about understanding statistical testing.

### **Studies on Foundations of Statistical Inference**

In this section, the research is reviewed with regard to methodologies used and major findings.

**Methodologies.** Most studies on people's understanding foundational ideas of statistical inference conducted are one-group posttest only evaluations with some variations in terms of settings, subject levels, sample size, and tasks examined.

First of all, most studies have been carried out in observational classroom settings (e.g., Carver, 2006; Lunsford, Rowell, & Goodson-Espy, 2006; Well, Pollastek, & Boyce, 1990, Study 1), with a few exceptions that included controlled conditions (e.g., Well et al., 1990, Study 3). Second, some studies included a specific course as a treatment (e.g., Konold, Pollastek, Well, & Lohmeier, & Lipson, 1993; Konold, 1994), but not always (e.g., Haller & Krauss, 2002). Third, researchers have used different methods for data collection—interviews (e.g., delMas & Liu, 2005; Kaplan, 2009; Konold et al., 1993) and a mixture of multiple-choice and open-ended questions (e.g., Haller & Krauss, 2002). Some researchers have used large-scale assessments (e.g., Carver, 2006; delMas & Liu, 2005, delMas, Garfield, Ooms, & Chance, 2006). They also

have also used think-aloud problem-solving protocols with a small number of questionnaires (e.g., Hertwig & Gigerenzer, 1999). Another factor that varies studies with regard to methodology is sample size—some studies have included a very small number of subjects (e.g.,  $n < 20$  in Hertwig & Gigerenzer, 1999), while others have employed larger sample sizes (e.g.,  $n > 50$  in Konold, 1994).

Age levels of the subjects range from primary students to undergraduates and teachers. Sample sizes range from small (e.g.,  $n = 10$  in Kaplan, 2009) to large numbers of subjects (e.g.,  $n = 114$  in a pre- and post-tests in Chance et al., 2004). A summary of the studies' characteristics (research design, sample size, subjects' grade level, and data collection methods) is presented in the Appendix A.

**Findings.** Most research on the topic of statistical inference has evolved from the early work of Daniel Kahneman and Amos Tversky. From studies about common human errors using heuristics and biases (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974), they established a cognitive basis for common human errors. They began their study with a detailed account of the *representativeness heuristic*, a tendency to assume that a sample represents the population regardless of its size. The following is a description of this heuristic shown in one of their instrumental papers, *Belief in the Law of Small Numbers* (Tversky & Kahneman, 1971):

People view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. Consequently, they expect any two samples drawn from a particular population to be more similar to one another and to the

population than sampling theory predicts, at least for small samples. (p. 24)

Kahneman and Tversky (1972) found this heuristic from university students and identified it as the *law of small numbers*, as opposed to the *law of large numbers*. Since their work, researchers have shown similar findings that people in general tend to look at a sample, just as a small part of a whole, and they place an excessive amount of confidence in small samples (e.g., Rubin, Bruce, & Tenny, 1991).

In a later study, Kahneman and Tversky (1982) conjectured that people tend to focus their attention to individual samples ignoring distributional propensities of the samples when making judgments under uncertainty. Compelling evidence for this conjecture was presented by Konold (1989). Referred to as the *outcome approach*, Konold found that people tend to base predictions of uncertain *individual* outcomes on causal explanations instead of on information obtained from repeating an experiment.

In a study by Rubin et al. (1991) with senior high school students, the researchers found that students have inconsistent models of the relationship between samples and populations. They also found that, for students who lack experience in thinking about a distribution of samples generated from a particular population, it is not easy to understand that “sample variability is the contrasting idea that samples from a single population are not all the same and thus do not all match the population” (p. 314).

In understanding of the idea of sample representativeness and sampling variability, the concept of sample size becomes important. Several studies have shown that students appear to have difficulty taking into account sample size in association with sample distributions (Mokros & Russell, 1995; Sedlemeier & Gigerenzer, 1997; Tversky

& Kahneman, 1971; Vanhoof, Sotos, Onghena, & Verschaffel, 2007; Schwartz, Goldman, Vye, Barron, & The Cognition and Technology Group at Vanderbilt, 1998).

Well et al. (1990) investigated how undergraduate students incorporate the information of sample size in sampling distributions. Four experiments differentiating different aspects of the problem revealed that students have incomplete conceptions of sample size. People appear to understand that the means of larger samples are more likely to resemble the population mean, but do not understand the implications of this fact for the variability of the sample mean, neglecting the effect of sample size in interpreting sampling variability.

Sedlemeier and Gigerenzer (1997) investigated 46 university students' understanding about frequency distributions and sampling distributions. They found that students did better at solving frequency distribution tasks than sampling distribution tasks, even when the participants fully understood the concepts given in sampling distribution tasks. Sedlemeier and Gigerenzer (1997) noted that students' intuitions regarding the empirical law of large numbers apply directly to frequency distributions, but not to sampling distributions. As a plausible reason for this, the researchers suggested, whether objects or events, the units of frequency distributions can be experienced in daily life, whereas proportions and means, the units of sampling distributions, are rarely experienced directly in everyday life.

Saldanha and Thompson (2002) conducted teaching experiments with senior high school students, and they found that students tend to focus on individual samples and statistical summaries of individual samples instead of looking at how collections of sample statistics are distributed. Saldanha and Thompson also found that students showed

a tendency to predict sample outcomes based on causal analyses instead of statistical patterns in a collection of sample outcomes. This finding is similar to the outcome approach studied by Konold (1989).

Thompson (2004) examined students' difficulty in understanding concepts of sampling distributions incorporating three major concepts: representativeness, variability, and sample size. He found that those students who seemed to understand the ideas and who used a margin of error for a sample statistic had developed what he called a "multiplicative conception of sample" (MCS)—a conception of sample that entailed variability among samples, the idea that each sample has an associated statistic that varies as samples vary. They argue that MCS enables students to understand the relationship between individual sample outcomes and distributions of a class of similar outcomes. In the same way, Saldanha and Thompson (2002) address that MCS empowers students to consider a sampling outcome's relative unusualness.

In terms of the misunderstandings students exhibit in reasoning about sampling distributions, results from studies tend to be consistent with the findings from studies.

These include:

- Students believe that the sampling distribution of a statistic should have the same shape and properties as the population distribution, indicating that students are confused about the population and the sampling distributions (e.g., delMas, Garfield, Chance et al., 1999a; 1999b).
- Students do not differentiate between the distribution of a sample and the sampling distribution of a statistic (e.g., mean; e.g., Lipson, 2003; delMas et al., 1999a; 1999b).

- Students do not understand the idea of the *law of large numbers* (delMas et al., 1999a; 1999b; Innabi, 1999).
- Students show misconceptions in understanding the concepts involved in the Central Limit Theorem (Batanero et al., 2004; Chance et al., 2004; delMas et al., 1999a, 1999b; Earley, 2001; Lunsford et al., 2006; Pfaff & Weinberg, 2009).

delMas, Chance, and Garfield (Chance et al., 2004) examined *how* they develop students' reasoning and *in what ways* instruction could help build students' inferential reasoning. The researchers designed five studies to investigate about difficulties students experience when learning about sampling distributions. In the fourth study, Chance et al. (2004) interviewed college students to gain a more in-depth understanding about students' conceptions of sampling distribution as well how they actually develop reasoning about sampling distributions. From the findings, they developed a framework to describe the development of students' statistical reasoning about sampling distributions, based on the work of Jones and colleagues (Jones, Thornton, Langrall, Mooney, Perry, & Putt, 2000). This framework consists of the following five levels of reasoning:

- Level 1—Idiosyncratic Reasoning: The student knows words and symbols related to sampling distributions, uses them without fully understanding them, often incorrectly, and may use them simultaneously with unrelated information.
- Level 2—Verbal Reasoning: The student has a verbal understanding of sampling distributions and the implications of the Central Limit Theorem,



but cannot apply this to the actual behavior of sample means in repeated samples.

- Level 3—Transitional Reasoning: The student is able to correctly identify one or two characteristics of the sampling process without fully integrating these characteristics.
- Level 4—Procedural Reasoning: The student is able to correctly identify the three characteristics of the sampling process, but does not fully integrate them or understand the predictable long-term process.
- Level 5—Integrated Process Reasoning: The student has a complete understanding of the process of sampling and sampling distribution, in which rules and stochastic behavior are coordinated.

As seen in a study by Chance et al. (2004), use of simulation in research studies is not rare in studies in the topic of sampling distributions. Incorporating simulation in the curriculum by using either hands-on activity (e.g., Chance et al., 2004; Pfaff & Weinberg, 2009) or computer software (delMas et al., 1999a; 1999b; Earley, 2001; Lane & Tang, 2000; Lipson, Kokonis, & Francis, 2003; Lunsford et al., 2006), researchers have investigated its impact on students' learning of sampling distribution concepts and analyzed in what specific areas students encounter difficulties.

Lane and Tang (2000) studied the effectiveness of simulations for teaching statistical concepts, compared to the effectiveness of a textbook. One hundred and fifteen undergraduate students were randomly assigned to the conditions of a factorial combination of “Medium” (computer simulation versus textbook) and “Question” (Specific versus Non-specific). This study revealed that training by simulation led to

better performance than training by a traditional textbook approach. The researchers found that simulation was especially effective when coupled with questions that focused students' attention to the relevant features or characteristics of the simulation.

Contrary to the results found by Lane and Tang (2000), several studies have revealed that simulation is not a sufficient way for students to develop reasoning of sampling distributions (delMas et al., 1999a, 1999b; Lipson, 2003, Lipson et al., 2003; Lunsford et al., 2006; Vanhoof et al., 2007). In delMas et al. (1999a; 1999b), researchers developed a computer simulation, Sampling Distribution, to facilitate students' learning of the concepts and ideas of the sampling process and distributions of samples. The researchers found that several students still did not appear to develop correct reasoning about sampling distributions, although there were some positive changes. Recognizing that simply showing students sampling distributions that are produced from random sampling does not improve students' understanding, in the next study, the researchers had the students make conjectures first. Based on their predictions about different empirical sampling distributions from various populations, students were then provided correct distributions. As a result, students' performance improved on the posttest when they were required to confront their misconceptions directly (delMas et al., 1999a; 1999b).

Lunsford et al. (2006) replicated the study of delMas et al. using the same conditions (post-calculus introductory course, use of the same assessment, software, and interview), but adding a pre- and post-survey to ask about students' reactions to specific instructional strategies. They found similar results to the previous researchers': Many students still showed incomplete reasoning, specifically in reasoning about the Central

Limit Theorem, although they showed improvement in post-tests after experiencing the computer simulation activity.

In summary, a number of studies provide many substantial works in research about people's understanding of sampling distributions. Observational studies with large sample sizes in some research studies have provided robust findings in terms of misunderstandings of the concept of sampling distributions. Teaching experiments and the use of various qualitative data have provided a framework to understand how students develop their reasoning about sampling distributions. Research studies on student understanding of sampling distributions have tended to employ both quantitative and qualitative methods. With relatively large sample sizes, many studies have revealed robust findings. There are also studies that examined why students encounter difficulty, and in what ways instruction may be helpful in improving their reasoning beyond identifying the misconceptions. There are also researchers who have examined the impact of simulation using hands-on activity or computer software.

### **Studies about Formal Statistical Inference**

Researchers agree that getting students to make sense of formal concepts and ideas in statistical inference is a very difficult goal for statistics instructors because of the persistence and deepness of misunderstandings held by learners (Daniel, 1998; Batanero, 2000; Sotos et al., 2007). Although educators recognize that students struggle with understanding formal statistical inference—the concepts and the logic of hypothesis testing, empirical studies are sparse on this topic compared to studies on sampling distributions. Characteristics of research studies on this topic are described next.

**Methodologies.** Most of the studies on reasoning about formal statistical inference present only a one-group posttest evaluation with some exceptions: implementing pre- and posttest (e.g., Falk & Greenbaum, 1995) or tests at multiple times (e.g., Pfaff & Weinberg, 2009). Only a few studies have used control conditions (e.g., Lane & Tang, 2000), while most are observational. The number of subjects has varied, from a small sample (e.g., 10 subjects in Kaplan, 2009) to a large sample (e.g., 436 subjects in Vallecillos, 2002). Although subjects are mostly college students, there are also some studies conducted with teachers (e.g., Haller & Krauss, 2002) or researchers (Mittag & Thompson, 2000). No studies were found that included subjects in primary or secondary school, supposedly because of the level of the topic. With regard to the methods of data collection, interviews (e.g., Williams, 1999a; 1999b), a mixture of multiple-choice and open-ended questions (e.g., Vallecillos, 1999), and surveys (e.g., Mittag & Thompson, 2000) have been used.

**Findings.** Although there are limited research studies on students' understanding of formal statistical inference such as hypothesis testing, researchers attempted to find difficulties and misunderstandings that students tend to show in learning the concepts of hypothesis tests. One of the studies found is by Liu and Thompson (2009). The researchers conducted a teaching experiment during professional development seminar. From the interviews with eight high school statistics teachers, the researchers identified the difficulties and conceptual obstacles that teachers experience in reasoning about the logic of hypothesis testing. The majority of the teachers failed to conceptualize a process of entailing a correct interpretation of unusualness. As an example, the following question was presented to the teachers:

Ephram works at a theater, taking tickets for one movie per night at a theater that holds 250 people. The town has 30,000 people. He estimates that he knows 300 of them by name. Ephram noticed that he often saw at least two people he knew. Is it in fact unusual that Ephram knows at least two people who attend the movie he shows? (p. 10)

Teachers' first responses to this question were mostly intuitive, such as, "It would not be unusual." In subsequent discussions, only one teacher had a conception of unusualness that was grounded in an understanding of the distribution of sample statistics. Other teachers have shown various conceptions of unusual; none of their reasoning is conceptualized based on repeated sampling that allows them to quantify unusualness. From this study, Liu and Thompson (2009) concluded that teachers' incomplete conceptions of probability results is a challenge when trying to understand inferences in hypothesis testing.

In addition to the conceptual challenges under the logic of hypothesis testing, Liu and Thompson (2009) found that teachers had difficulty in conceiving the role of hypothesis testing as a tool for making a conclusion from inferences. In their study, teachers appeared not to internalize the functionality of hypothesis testing, showing a lack of understanding as to how hypothesis testing can be a useful tool for making decisions.

Vallecillos (2002) found similar results from university students. Examining 436 university students' understanding of hypothesis testing, he found that students do not consider hypothesis testing as a process of decision making to accept or reject a hypothesis. Vallecillos identified four different conceptions regarding the type of proof

that hypothesis tests provide: a) conception of the test as a decision-making rule; b) conception of the test as a procedure for obtaining empirical support for the hypothesis being researched; c) conception for the test as a probabilistic proof of the hypotheses; and d) conception of the test as a mathematical proof of the hypothesis' truth.

Confusion about the logic of hypothesis testing was also shown in a study by Williams (1999a; 1999b). Conducting interviews with 18 students in an introductory statistics course, he investigated about students' conceptual and procedural knowledge of significance level. A concept map was used to assess students' conceptual knowledge, and formal hypothesis test tasks were used to assess procedural knowledge. Students were asked to talk aloud as they completed a concept map task and two formal hypothesis-testing tasks. On the conceptual test, students' understandings about the definition of significance level varied from seeing it as representing a level for decision-making, a measure of significance, or a level of confidence or error. On the procedural test, students demonstrated a confusion between *P*-values and significance level.

Smith (2008) examined existing differences in students' understanding between the concepts and procedures of hypothesis testing. In order to explore how undergraduate students develop an overall "big picture" of statistical hypothesis testing, she examined 104 introductory students' understanding of hypothesis testing using a 14-item multiple-choice questionnaire. She also conducted follow-up interviews with 11 students who presented a range of performance patterns on the questionnaire. In this study, Smith found that students did not have high degrees of conceptual understanding or adaptive reasoning. Although students were able to perform the procedures, students did not have

strong understandings of the concepts, logic, and uses of the methods in hypothesis testing.

While these researchers examined students' understanding of methods of formal statistical inference as an entire process (e.g., process of decision making, or logic of hypothesis testing), some researchers have focused on specific topics involved in formal statistical inference (e.g., the meaning of statistical significance,  $P$ -values, or the role of sample size in hypothesis tests). Wilkerson and Olson (1997) surveyed 52 graduate students to investigate about students' understanding of the relationships between treatment effect, sample size, and statistical significance. Results from the survey revealed that student responses placed more confidence in the results of studies with large sample sizes than in the results of studies with small sample sizes, regardless of the criterion on which that confidence was based. A significant number of respondents failed to recognize that a small sample requires a greater treatment effect than a large sample to obtain an equal level of statistical significance.

A study conducted by Haller and Krauss (2002) showed people's misunderstanding of significance tests and  $P$ -values. Methodology instructors, scientific psychologists, and psychology students in German universities were included as subjects. The researchers provided them with six true-false items representing "common illusions" of the meaning of a significant test result. In this study, many instructors and psychologists tended to show incorrect understanding about *how to interpret a significant result from hypothesis testing*. For those six statements of interpreting a significant result, nearly 90% of psychologists and 80% of methodology instructors showed at least one of

the false “meanings” of a  $p$ -value (e.g., “You have found the probability of null hypothesis being true”).

From researchers’ effort to find empirical evidence of what specific misunderstandings occur, specific areas that people showed difficulties in understanding of formal statistical inference are identified in the literature and listed below:

- The definition of the hypotheses (e.g. Vallecillos & Batanero, 1997)
- The nature (role) of hypothesis tests (e.g., Mittag & Thompson, 2000)
- The conditional logic of significance tests (e.g., Haller & Krauss, 2002)
- The interpretation of  $P$ -values (e.g., Williams, 1999a; 1999b)
- The evaluation and interpretation of statistical significance (e.g., Wilkerson & Olson, 1997)

Although there seems to be an agreement on what misconceptions people show in formal statistical inference, there is little empirical research about where these misconceptions come from and how to improve students’ understanding of the concepts in hypothesis testing. A research study by Kaplan (2009) provides a possible explanation of why students show difficulty in inferences of hypothesis testing. She conducted a study about grounded conception, which prevents sound reasoning. She specially focused on the impact of “Belief Bias” discovered by psychologists, which is a tendency “to rate the strength of arguments based on the believability of the conclusions” (Kaplan, 2009). In interviews with ten undergraduate students, she asked about three scenarios, varying in degrees of believability (low, moderate, and high believability). Each task included a description of an experimental study with statistical conclusions, along with  $P$ -values and interpretations of the results of the hypothesis test. In the given tasks, students showed



three types of evidence as being convincing: 1) statistical results; 2) a preponderance of evidence; and 3) a justification or rationalization. In addition, students tended to be less convinced by the statistics when the conclusion suggested by statistical evidence was incongruent to a prior belief. In this case, students tried to search for a justification of the conclusion, or they relied on their preexisting opinions.

Although Kaplan's study provided one plausible explanation of students' difficulty in understanding inferences involved in formal statistical inference, the sample size and type of tasks limit the generalization of results of this study to a larger context. Studies of other factors that could also influence people's misunderstanding of formal statistical inference were not found in the literature.

### **What is Informal Inferential Reasoning (IIR)?**

Given that students show consistent difficulties in understanding and reasoning about formal statistical inference, researchers and educators have been trying to find ways to develop students reasoning about statistical inference. One of the attempts is to expose them to situations where they use informal reasoning. Garfield and Ben-Zvi (2008) suggest that ideas of inference should be introduced *informally* at the beginning of the course, such as having students become familiar with seeing where a sample corresponds to a distribution of sample statistics, based on a theory or hypothesis. They further argue that this may help students be less confused by the formal ideas, procedures, and language when they finally reach the formal study of this topic.

Ben-Zvi (2006) also argues that statistical inference is essentially informal, although teaching inference in statistics has focused on formal methods. Similarly, Pratt et al. (2008) maintain that conceptual struggle in statistics needs to take place for students

in order to engage in informal inferential reasoning from a constructivist stance. As noticed, these researchers consider informal inference as a way to support a coherent understanding of formal concepts in statistical inference. In addition, it appears that they are more interested in students' naïve conceptions than identifying of students' misunderstandings of reasoning about formal statistical inference.

In the next section, the terms that are used about statistical inference are clarified. The section also presents definitions, describes characteristics of Informal Inferential Reasoning (IIR) and reviews research studies on IIR.

### **Inferential Reasoning in Statistics (IRS)**

In literature about statistical inference, it appears that different terms are used interchangeably (e.g., statistical inference, inferential reasoning in statistics, and reasoning about statistical inference). Specifically, research literature seems to use the two terms without distinguishing between *statistical inference* and *reasoning about statistical inference*. For instance, in Sotos et al. (2007), the researchers use the term statistical inference as a *content domain* that includes several topics in it (e.g., “a core idea in the understanding the concepts in statistical inference”). However, in Zieffler et al. (2008) statistical inference refers to a *reasoning process* (e.g., “formal methods of statistical inference”).

To clarify the uses of the terms, this study refers to the term statistical inference as *a content domain* that involves the concepts and ideas related to inferential statistics. As reviewed in previous sections, this includes foundations of statistical inference (e.g., sampling distribution) and formal statistical inference (e.g., hypothesis testing). Statistical inference also includes the topics regarding informal inference, which is described in this

section. Differentiating from statistical inference as a content domain, this study uses the term inferential reasoning in statistics (IRS) as *reasoning* that people use to understand the concepts and ideas of statistical inference. IRS is defined as the way people draw conclusions from data at hand to a broader context using the concepts and ideas of statistical inference. The relationships between the terms are illustrated in Figure 1.

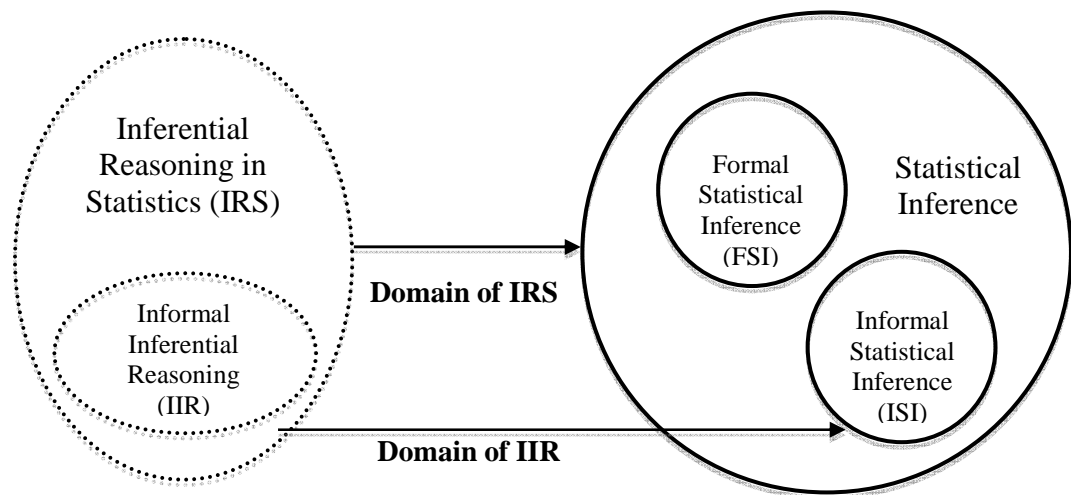


Figure 1. Conjectured relationships between the terms related to statistical inference.

### Definition and Components of IIR

Rubin, Hammerman, and Konold (2006) describe IIR as a construct formed with multiple dimensions. In this perspective, IIR has properties of “aggregated rather than properties of the individual cases themselves, signals and noise, various forms of variability, sample size, controlling for bias, and tendency” (p. 2). The multi-faceted aspect of informal inference is also shown in the definition of IIR suggested by Pfannkuch (2006b). She explains that IIR is the ability to interconnect ideas of— distribution, sampling, and center, within an empirical reasoning cycle. Makar and Rubin

(2009) also view informal inference as a multi-faceted construct, and provide a detailed description:

By formal statistical inference, we refer to inference statements used to make point or interval estimates of population parameters or formally test hypotheses, using a method that is accepted by the statistics and research community. Informal statistical inference is a reasoned but informal process of creating or testing generalizations from data, that is, not necessarily through standard statistical procedures. (p. 85)

Makar and Rubin's (2009) description about informal inference seems to include two key components: (1) making an inference about a population or testing hypotheses, and (2) a process of inference that does not utilize (formal) statistical procedures. These two components are also seen in Rossman's (2008) perspective on IIR where he describes IIR as "going beyond the data at hand" and "seeking to eliminate or quantify chance as an explanation for the observed data" through an argument with no formal method, technique, or calculation (as cited in Zieffler et al., 2008).

Ben-Zvi (2006) includes an argumentation component to this definition of IIR. He describes argumentation as a "discourse for persuasion, logical proof, and evidence-based belief, and more generally, discussion in which disagreements and reasoning are presented" (p. 2). From Toulmin's argumentation model (1958)—which consists of data, warrant, backing, qualifier, reservation and claim—Ben-Zvi notes that the integration and cultivation of informal inference and informal argumentation are essential in constructing students' statistical knowledge and reasoning in rich learning contexts.

Incorporating different perspectives on IIR including Makar and Rubin's (2009) and Ben-Zvi's (2006), Zieffler et al. (2008) provide a working definition of informal inference: "the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples" (p. 44). Zieffler et al. (2008) also provide components of an informal inferential reasoning framework. The components are:

- Making judgments, claims, or predictions about populations based on samples, but not using formal statistical procedures or methods (e.g., *P*-values, *t*-tests);
- Drawing on, utilizing, and integrating prior knowledge (e.g., formal knowledge about foundational concepts; informal knowledge about inference, such as recognition that a sample may be surprising, given a particular claim; use of statistical language), to the extent that this knowledge is available; and
- Articulating evidence-based arguments for judgments, claims, or predictions about populations based on samples. (p. 45)

In summary, IIR is described as the way that people reason using interconnected informal knowledge or ideas to make claims about population and to support inferences from observed samples to the population. IIR is differentiated from formal statistical reasoning in that, in IIR, standard statistical procedure or concepts (e.g., hypothesis tests, *p*-value, or statistical significant) are not necessarily used.

## **Role of IIR in Reasoning about Statistical Inference**

In general, informal reasoning is useful when information is less accessible, or when the problems are more open-ended, debatable, complex, or ill-structured, especially when the issue requires individuals to build an argument to support a claim (Means & Voss, 1996). In statistics education, IIR is considered as “a potential pathway” for supporting students’ understanding of formal statistical concepts (Makar and Rubin, 2009). A similar, but more detailed role of IIR is presented in Zieffler et al. (2008):

[Because] statistical inference integrates many important ideas in statistics—such as data representation, measures of center and variation, the normal distribution, and sampling—introducing informal inference early and revisiting the topic throughout a single course or curriculum across grades could provide students with multiple opportunities to build the conceptual framework needed to support inferential reasoning. (p. 46)

From this paragraph, it seems that the essential role of informal inference is that the IIR can be used to support students’ IRS as they learn important ideas in statistical inference.

Makar and Rubin (2009) also advocate that IIR provides “new opportunities to infuse powerful statistical concepts very early in the school curriculum and return the focus of statistics to a tool for insight into understanding problems rather than only a collection of graphs, calculations, and procedures” (p.102), a notion that has also been addressed by other researchers (e.g., Ben-Zvi & Sharett-Amir, 2005; Sorto, 2006).

## **Studies on Informal Statistical Inference**

Given the role of IIR in reasoning in statistical inference, educators and researchers have attempted to define and characterize IIR. There have been also some empirical research studies on this topic.

Pfannkuch (2005) conducted a case study in a grade 10 classroom after the teacher participated in a workshop to investigate about statistical thinking of teachers as well as students. The subjects' attempts at informal inference with boxplots were examined from student bookwork, student responses to assessment tasks, and the teacher's weekly audiotaped reflections, the researcher investigated students' attempts at informal inference with boxplots. From an analysis of students' responses, Pfannkuch found that students did not tend to explain how their analyses supported their conclusions even though their responses were appropriate in relation to the question and they drew a valid conclusion in the comparison of data sets. Pfannkuch (2005) proposed two conjectures as possible explanations for this result: (1) the current curriculum tended to compare only the features of boxplots and not drawing conclusions; and (2) the curriculum did not provide a teaching pathway to build students' concepts of formal inference, nor did it provide learning experiences for the transition between informal and formal inferential thinking.

Pfannkuch (2005) suggests a framework for developing the concepts of informal inference that includes—reasoning with measures of center, distributional reasoning, sampling reasoning, and drawing an acceptable conclusion, based on informal inference. Raising further questions as to what types of learning experiences would develop students' inferential reasoning toward a more formal level, Pfannkuch conducted a larger

project (2006a; 2006b; 2007). Using an action research approach, Pfannkuch (2007) compared a teacher's reasoning to students' reasoning when drawing informal inferences from a comparison of boxplots. From a qualitative analysis of the teacher's communication to her students during three teaching episodes, Pfannkuch (2007) extracted eight descriptors of informal reasoning—hypothesis generation, summary, shift, signal, spread, sampling, explanatory, and individual case. She found that the teacher's view of the inferential task was multifaceted and incorporated all of the eight descriptors. Using the same descriptors, Pfannkuch (2007) analyzed students' reasoning on the same tasks. Of 26 students, only 11 were reasoning beyond a *descriptive* view, and the *sampling* view was not present in the students' responses. This finding indicates that the students found it difficult to verbally express, describe, and justify conclusions when comparing boxplots. Pfannkuch (2007) argues that the students were not given opportunities to have experiences involving sampling variability or sample size effects. She further argues that in order to develop students' inferential reasoning from distributions, instruction needs to address and build concepts about sampling behavior.

The results of students' incomplete understanding about the boxplot comparison are consistent to a study by Biehler (2005). He found that students tended to reason with and compare five-number summary cut-off points when dealing with boxplots without considering the spread. He also found that students did not exhibit a shift view, where the majority of the data appears to shift positions from one dataset to another, nor did they have intuitions about sampling variability.

Makar and Rubin (2009) developed a model to characterize informal statistical inference. They investigated the thought processes of primary schoolteachers' learning in



teaching mathematics and statistics through inquiry in a problem-based environment. The subjects were four primary schoolteachers in Australia. Using data from videotapes of the teachers' lessons, collections of lesson plans, student work, and interviews of the teachers, the researchers examined how teachers teach informal inferential reasoning. Three principles of informal inferential reasoning were proposed: generalizations beyond the data, data as evidence, and probabilistic language.

Using a design experiment, the authors further investigated this framework to consider the way that students and teachers could employ inferential reasoning when working with data. In terms of *generalization*, the researchers found three elements missing from teachers' descriptions of what they considered to be important in providing opportunities for students to tap into inferential reasoning: pose a driving question; include an engaging context; and ensure sufficient complexity in the data.

With regard to *data as evidence*, the teachers focused on making generalizations from the data, which supported students in seeing the data as evidence for their conclusions. However, students' attention to descriptive statistics (e.g., graphing skills) never got back to the problem, which would have allowed them to make the connection between the data they collected and their potential as evidence for drawing inferences. From this result, the researchers found that the use of data as evidence is a key principle of informal inference that reminds learners of: (1) the purpose of collecting and analyzing data; and (2) the importance of focusing on the problem and process of statistics in inquiry rather than merely a dataset as an isolated artifact.

The third principle of informal statistical inference, *probabilistic language*, appeared to be the most apparent aspect of informal inference. In the context that students

used the data they had collected on handspans, students' language changed to include notions of uncertainty and level of confidence once they made the connection between using their own data as evidence to make predictions. Articulating students' uncertainty in making predictions allowed students to take a risk without worrying about possibly being "wrong" by using notions of uncertainty and levels of confidence. Encouraging the students' ability to articulate their uncertainty by using their own dataset as evidence to make predictions can be a way to enhance students' ability to express, describe, and justify their reasoning, which were shown to be difficulties in the study by Pfannkuch (2006a; 2006b).

Another substantial study about students' informal reasoning is a study by Ben-Zvi (2006). Using developmental research, he investigated the emergence of fifth-grade students' informal reasoning. Students' learning processes were analyzed as they learned the *growing samples* instructional heuristic (Bakker & Gravemeijer, 2004) with the software *TinkerPlots*. From an analysis of the videotapes, observations, and interviews of selected students and teachers, he identified levels of changes in students' statistical reasoning in multiple dimensions: progress from additive to multiplicative reasoning; consideration of aggregate views of data; acknowledgement of the important role of larger samples; and accounting for variability. He found that "the emergence of students' statistical knowledge was accompanied by the growing ability to discuss their thoughts and actions, explain their inferences and argue about data-based claims" (p. 5).

In terms of the factors that influence the development of students' informal statistical reasoning, Ben-Zvi and Gil (2010) investigated the role of context in the setting of extended curriculum development with three sixth-grade students. They found that

context played a role of resolving conflicts between expectations and data by helping break through unclear or contradicting points in understanding graphs.

In sum, research studies on the topic of informal statistical inference have identified the role of IIR in association with IRS, the components of IIR, and the role of context in IIR. While these studies offered information about fundamentals agreed upon by researchers, there are few empirical research studies, specifically, on the issues of—how to improve students' informal inferential reasoning, how students' actual IIR relates to their IRS, and what instructional methods are effective in helping students to support IRS using their IIR.

### **Content Domains of Statistical Inference**

Research studies and literature reviewed on the topic of IIR suggest that the content domain of IIR may be represented by two content areas: informal statistical inference (ISI) and formal statistical inference (FSI). These categories are used as the content domain of IRS, and thus, cover the contents of statistical inference. The contents of based on the literature review are listed below.

- The concept of uncertainty (Makar and Rubin, 2009)
- Properties of aggregates (Makar and Rubin, 2009; Pfannkuch, 1999; Rubin et al., 2006)
- The concept of sampling variability (Rubin, Hammerman & Konold, 2006; Pfannkuch, 1999; Wild et al., 2011; Zieffler et al., 2008)
- The concept of unusualness (Liu and Thompson, 2009; Makar and Rubin, 2009; Rubin et al., 2006; Zieffler et al., 2008)
- Generalizing from a sample to a population (Zieffler et al., 2008)

- Comparison of two populations from two samples (Makar and Rubin, 2009; Pfannkuch, 2005; Wild et al., 2011; Zieffler et al., 2008)

FSI includes the methods and concepts used in formal inferential statistics. In addition, FSI also includes the fundamental concepts in formal statistical inference such as sampling distribution considering that those concepts represent an important building block to a coherent understanding of statistical inference. Thus, the topics of FSI have been identified following the same structure as in the literature review—foundations of statistical inference and formal statistical inference. Following are the foundations of statistical inference:

- The concepts of samples and sampling (Saldanha and Thompson, 2002; Saldanha, 2004; Rubin et al., 1991)
- Law of Large Numbers (Sample representativeness; Kahneman and Tversky, 1972; Metz, 1999; Rubin et al., 1991; Saldanha & Thompson, 2002; Watson & Moritz, 2000)
- Population distribution and frequency distributions (delMas et al., 1999a, 1999b; Lipson, 2003)
- Population distribution and sampling distributions (delMas et al., 1999a, 1999b)
- Central Limit Theorem (Mokros and Russell, 1995; Sedlemeier & Gigerenzer, 1997; Tversky & Kahneman, 1974; Schwartz et al., 1998; Vanhoof et al., 2007; Wagner & Gal, 1991; Well, Pollastek, and Boyce, 1990)

Formal statistical inference:

- Definition, role, and logic of hypothesis testing (Batanero, 2000; Haller & Krauss, 2002; Liu & Thompson, 2009; Mittag & Thompson, 2000; Nickerson, 2000; Vallecillos, 2002; Williams, 1999a, 1999b)
- Definitions of  $P$ -value and statistical significance (Carver, 1978; Falk & Greenbaum, 1995; Nickerson, 2000)
- $P$ -value as a numerical probability (Cohen, 1994; Rosenthal, 1993)
- Sample size and statistical significance in hypothesis testing (Wilkerson & Olson, 1997)
- Confidence interval (Fidler, Thomason, Cumming, Finch, & Leeman, 2004)
- Evaluation of hypothesis testing (Wilkerson & Olson, 1997)

### **Need for an Instrument to Assess Inferential Reasoning in Statistics**

#### **New Instructional Approaches to Develop Students' Understanding of Statistical Inference**

Along with ongoing calls for reform in introductory statistics courses, different teaching methods for developing students' inferential ideas have been proposed. Cobb (2007) and Kaplan (2007) have suggested a radical approach to statistical inference in the introductory course. Cobb (2007) argues that statistics educators need to reconsider both the pedagogy and the content of introductory statistics courses in that the approach from asymptotic sampling distributions centered around the normal distribution turned out to be cognitively complicated. Addressing that the logic of inference should be at the center of introductory course instead of the normal distribution, he argues that a randomization approach (e.g., permutation tests) should be used as a main method to teach statistical

inference. He further addresses that using permutation tests to teach statistical inference provides students with both conceptually easier instruction for statistical inference and a modern, computational data analysis technique currently lacking in the first course in statistics.

Inspired by Cobb, recent NSF-sponsored projects have developed new curriculum to introduce students to ideas of statistical inference using randomization methods (The CSI project, headed by Rossman, Chance, Cobb, & Holcomb (<http://statweb.calpoly.edu/csi>); The CATALST course, developed by Garfield, delMas and Zieffler (<http://www.tc.umn.edu/~catalst>); The INCIST project, headed by West and Woodard). Given the current interest in randomization-simulation methods that are being currently implemented in some statistics courses, several questions need to be addressed: What is the impact of those curricula on students' inferential reasoning? How do the students taught with statistics curricula based on randomization-simulation approaches differ from the students taught with traditional curricula (based on asymptotic sampling distributions)? How do we know how students are doing in these courses? In order to address these issues, there is a need for a research instrument to assess students' outcomes with regard to this innovative approach.

### **Existing Assessments**

There have been some studies on the development of assessments in statistics targeting college students. The Statistical Reasoning Assessment (SRA; Garfield, 2003) was designed to assess students' ability to reason with statistical information (e.g., correctly interpreting probability, understanding independence and sampling variability, distinguishing between correlation and causation). The SRA has been used in different

contexts, and reasonable test-retest reliability and content validity have been established (Garfield, 1998b, 2003; Liu, 1998). However, it focuses heavily on probability and lacks items related to data production, data collection, and statistical inference (Garfield, 2003).

The Statistics Concepts Inventory (SCI) was developed to assess statistical understanding, but it was written for a specific audience of engineering students in statistics (Reed-Rhoads, Murphy, & Terry, 2006). The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project was designed to develop an assessment instrument that would have broader coverage of both the statistical content typically covered in beginning, non-mathematical statistics courses, and would apply to the broader range of students who enroll in these courses (Garfield et al., 2002). This result was the Comprehensive Assessment of Outcomes in a First Statistics course (CAOS, delMas et al., 2007). The CAOS is a 40-item test that was designed to evaluate student attainment of desired outcomes in an introductory statistics course. The items focus on the big ideas and “the types of reasoning, thinking and literacy skills deemed important for all students across first courses in statistics” (Garfield et al., 2002).

### **Why These Assessments Do Not Meet the Current Need**

While these instruments share some characteristics, such as assessing students reasoning, thinking, and conceptual understanding rather than procedural skills of problem, the existing instruments are not appropriate to measure students' IRS. The existing instruments do not measure the full domains of inferential reasoning in statistics. Moreover, these existing instruments do not align with the use of randomization-simulation methods in terms of the contents of a test designed specifically to target developing students' reasoning about statistical inference. In addition, all of these

assessments are outcome-based assessments with formats similar to proficiency or achievement tests, which have limitations for providing educators with information for formative assessment.

Another limitation of the existing instruments is that they were not developed or validated using psychometric measurement models (e.g., item response theory) that provide ample information about properties of items (e.g., item difficulty, item discrimination, item information). Therefore, there is a need for a new instrument that will assess the content areas of informal and formal inference, be well-aligned with the new randomization-simulation based curriculum, and be developed and validated using appropriate psychometric theory.

### **Summary of the Literature Reviewed**

A summary of the methods used in studies of statistical inference in different fields appears in Appendix A. Findings shown in research studies have suggested that many students taking introductory statistics courses do not seem to understand much of what they are studying. Students encounter challenges when they learn the formal processes and concepts in inferential statistics. Studies have documented many of these challenges and have tried to uncover the reasons *why* people have difficulty with statistical inference. Kahneman and Tversky's early works on this topic have contributed to the literature of characteristics regarding people's judgment under uncertainty. Studies have also revealed and identified common misconceptions people make in statistical inference, such as the representativeness heuristics, the law of small numbers, and misconceptions regarding *P*-values or the logic of hypothesis testing. Researchers have



tried to provide a framework to guide understanding of student's development of reasoning about statistical inference.

These findings have many implications for the teaching and assessing statistical inference. One such implication is in the curriculum, which is covered in the classroom. While current curriculum documents (e.g., NCTM 2000 and GAISE reports) provide suggestions of teaching concepts of statistical inference, research strongly suggests (e.g., Chance et al., 2004) that large numbers of students fail to comprehend formal statistical inference when they meet it in introductory statistics courses. Recent research reviews have pointed to the importance of building up “the staged development of the big ideas of statistical inference” (e.g., Wild et al., 2011, p.1) rather than presenting formal concepts directly. One way of building the big ideas of statistical inference suggested from research studies is to have students begin working with precursor forms of statistical inference. This idea is congruent with recommendations by Ben-Zvi (2006), Pfannkuch (2005, 2006a, 2006b), Makar and Rubin (2009) and Zieffler et al. (2008).

Another implication concerns the use of technology in teaching statistical inference. The use of simulation to explore sampling distribution and hypothesis testing has shown that students can better capture the behaviors of sample statistics through a dynamic visual approach. Among many benefits of this approach is that technologies can create multiple and linked representations (e.g., boxplots of two datasets), and thus, it allows students to make a decision about whether one group is bigger than another by providing a big picture *before* using formal methods (e.g. t-test or permutation test).

## **Formulation of the Problem Statement**

Despite the influential contribution of the works, the studies that have been reviewed on statistical inference leave room for many new studies and research questions. For example, most of the studies used qualitative methods in nature. Although many of the qualitative studies provided substantial findings by examining subjects closely, there is a lack of quantitative evidence that could better answer some questions, such as which instructional methods in teaching formal concepts in statistical inference will improve students' understanding of the ideas of statistical inference.

A few studies have employed quantitative methods with large sample sizes. However, most of the quantitative studies used observational data with only a one-group posttest or quasi-experimental design with no randomization. The samples employed have usually been convenience samples. In addition, the instruments used to examine students' reasoning have not been validated in terms of psychometric properties, such as reliability, validity, or discrimination.

Most of the literature on students' learning of inferential reasoning has examined partial aspects of statistical inference, such as, whether or not students can reason correctly for given specific questions or tasks. Many of the concepts of statistical inference in an introductory statistics course that students are expected to understand after taking the course are not explicitly addressed by the research. Studies are needed in areas where students show appropriate reasoning or misunderstanding in a systematic view in order to examine their inferential reasoning as an entire process.

More studies are needed to find out the extent of student understanding and misconceptions for a wide variety of statistical inference concepts, but beyond looking at

whether they understand some specific concepts or not. Considering statistical inference as the ability to think “beyond the data at hand and to draw conclusions about some wider universe by taking account that variation is everywhere and the conclusions are uncertain” (Moore, 2007, p.172), students’ reasoning about statistical inference can be better captured by examining how they reason and how they make a decision in a well-structured contextual frame.

Taken together, there are several questions that have not yet been answered in the literature: What is the impact of an instructional approach designed to develop students’ inferential reasoning? Is there any structure in statistical inference distinguishable by informal and formal inferences? How do these two types of statistical inference relate to each other? What would be the best way to measure these two types of inferential reasoning? These questions lead to the need of an instrument that measures students’ reasoning about statistical inference in multiple aspects as a whole, so that statistics educators could guide and monitor students’ developing ideas of statistical inference. With a reliable and valid measure, the questions listed above could be meaningfully investigated.

The research describes the development and the validation of an instrument to measure college students’ inferential reasoning in statistics. The research questions to be addressed are:

1. To what extent are the scores on the proposed test precise?
2. To what extent are the scores on the proposed test generalizable to a larger domain?

3. To what extent do the scores on the proposed test reflect students' actual reasoning in statistics?
4. To what extent do items reflect the structure of ISI and FSI?

## Chapter 3

### Methods

This chapter discusses the procedures for gathering and analyzing the data obtained in the study. The literature reviewed in the previous chapter suggests that there is a need to develop a new instrument to measure inferential reasoning in statistics (IRS) and that IRS be represented to two content categories—informal statistical inference (ISI) and formal statistical inference (FSI). In response to this need for a new instrument, this study developed and validated an assessment for measuring college students' IRS in two areas—ISI and FSI.

The argument-based approach to validity (Kane, 1992, 2006a, 2006b) was used as a theoretical framework to guide the process of test development and validation, which is described in the first section. The second section provides a framework of the study structured to formative stage and summative stage. Different sources of validity evidence gathered in each stage are described in the next section. Theoretical evidence obtained in formative stage is presented first. A description of empirical evidence collected in summative stage is followed. For each of the data sources are outlined in terms of the resources of data, participants and procedures of data collection. This section also explains the methods of data analysis including local item dependency (LID), dimensionality, and item response theory.

### **Validity and Validation**

#### **Validity**

Validity is the most fundamental consideration in developing and evaluating tests. According to the *Standards for Educational and Psychological Testing* (hereafter referred

to as *Testing Standards*; AERA, APA, NCME, 2002), validity “...refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). Messick's (1989) definition emphasizes the appropriateness of score-based actions in addition to the appropriateness of inferences:

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test scores or other modes of assessment. (p. 13)

Cronbach (1971) defines validity in terms of “the soundness of all the interpretations of a test.” All of the above definitions relate validity to the appropriateness of the inferences included in test score interpretations.

Sireci (2007) describes the fundamental aspects of validity, as follows:

- Validity is not a property of a test. Rather, it refers to the use of a test for a particular purpose.
- Evaluating the utility and appropriateness of a test for a particular purpose requires multiple sources of evidence.
- If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose.
- Evaluating test validity is not a static, one-time event; rather, it is a continuous process.

Sireci (2007) argued that an iterative process is necessary to evaluate the adequacy of test score interpretations from the proposed assessment. Therefore,

validation is, in itself, a process of collecting and accumulating multiple sources of evidence to evaluate inferences from test scores to various conclusions.

In the *Testing Standards* (AERA et al., 2002), it is stated that the process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. The conceptual framework points to the kinds of evidence to be collected in order to evaluate the proposed interpretation in light of the purposes of testing.

It is noted that different aspects of validity will be illuminated by various sources of evidence that will support validity as a unitary concept. As the *Testing Standards* notes, the different aspects of validity do not represent distinct types of validity. Rather, they represent diverse perspectives that are integrated to provide evidence that supports validity for the use of the proposed assessment. Following this suggestion, this study identifies each source of validity evidence according to the origin of the evidence. The sources of validity evidence identified in *Testing Standards* are described below.

Evidence based on test content is obtained from an analysis of the relationship between a test's content and the construct it is intended to measure. It is also obtained from a specification of the content domain. The evidence can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretations of test scores (AERA et al., 2002).

Evidence based on response processes comes from analyses of individual responses. Theoretical and empirical analyses of the response processes of test takers can

provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by examinees (AERA et al., 2002).

Evidence based on relationships with other variables indicates the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based. An analysis of the relationship of test scores to the variables external to the test provides another important source of validity evidence. According to the *Testing Standards* (AERA et al., 2002), external variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs.

Evidence based on the internal structure of a test addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretations. An estimate of score reliability or examination of differential item functioning is some examples of this validity evidence.

Another source of evidence described in *Testing Standards* is evidence based on consequences of testing. Evidence based on consequences of testing concerns an issue of incorporating the intended and unintended consequences of test use into the concept of validity. Claims are sometimes made for benefits of testing that go beyond direct uses of the test scores themselves (e.g., test uses to improve student motivation or changes in classroom instructional practices). The validations of such cases are then examined by evidence that the anticipated benefits of testing are being realized (AERA et al., 2002).



## **A Validation Method: An Argument-based Approach to Validation**

An argument-based approach to validation was suggested by Kane (1992, 2001, 2002) by building on the work of Cronbach (1971, 1988), House (1980), and Messick (1989). Kane (1992) argued that test-score interpretation is associated with a chain of interpretive arguments, and that the validity of interpretation and uses of the test-score are determined by the plausibility of those arguments. In this approach, interpretive arguments establish a network of inferences from observations to score-based conclusions and decisions, and guide the collection of relevant evidence that supports those inferences and assumptions. Therefore, validity is an argument construed by an analysis of theoretical and empirical evidence instead of a collection of separate quantitative or qualitative evidence (Bachman, 1990; Chapelle, Enright, Jamieson., 2008, 2010; Kane, 1992, 2001, 2002; Mislevy, 2003). In this sense, validity cannot be proved, but depends on the plausibility of interpretive arguments that can be critically evaluated with evidence.

From the previous works by Cronbach (1971, 1988), House (1980), and Messick (1989), Kane (1992) addressed the importance of making proposed interpretations and uses explicit through an interpretive argument. This interpretive argument specifies the inferences and assumptions leading from test scores to the interpretations and decisions based on test scores (Kane, 2006a). This interpretive argument is articulated through a validation process that considers the reasoning from the test score to the proposed interpretations and the plausibility of the associated inferences and assumptions. This set of inferences and assumptions are then evaluated by examining the validity argument developed from the interpretive argument. The different types of validity evidence are

gathered to support the validity argument as claims, intended inferences, and assumptions. In this process, four inferences provide the framework that encompasses each inferential link based on an assumption that must be evaluated:

1. Scoring: an inference from an observation of performance to a score.
2. Generalization: an inference from the observed score on a particular test to a universe score, which assumes that the observed score is based on random or representative samples from the universe of generalization.
3. Extrapolation: an inference from the universe score to a target score.
4. Explanation/Implication: an inference explained about the estimated target score regarding a description of knowledge, skills, or abilities.

The argument works if these inferences can be justified from validity evidence by addressing how convincingly the evidence supports the network of inferences. These inferential links in an interpretative argument are illustrated in Figure 2.

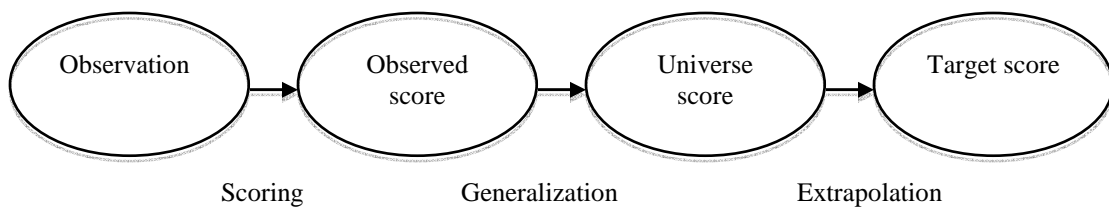


Figure 2. Kane, Crooks, & Cohen (1999).

This network of inferences forms the interpretive argument from an observed test performance to the conclusions; and the interpretive argument is examined in terms of

plausibility based on assumptions (e.g., theories, empirical generalizations, factual statements; Kane, 2006a).

### **Framework of the Study**

With the framework of an argument-based approach to validation, this study was structured to two stages: a formative and summative stage, as Kane (1992) proposed. In the formative stage, a test blueprint and preliminary assessment were developed from a review of the literature and expert reviews. The test domains and a set of tasks were specified. This stage also involved formulating the interpretative argument by clarifying the inferences and assumptions regarding test score interpretations and uses. A set of proposed claims regarding the test score uses were derived from the intended purpose of the assessment. Since the purpose of the assessment is to examine and provide information about students' current standing on IRS rather than to make decisions (e.g., placement or certification), interpretations of the test score were descriptive rather than decision-based or prescriptive (Kane, 2001, 2002). Therefore, this study involved descriptive interpretations regarding inferences from an observed score to a target score.

In the summative stage, a set of interpretive arguments was evaluated as the different sources of validity evidence were gathered. The evidence sources were identified based on *Testing Standards* (e.g., evidence based on contents). The interpretive argument was examined in terms of plausibility of the associated assumptions specified by four inferences (scoring, generalization, extrapolation, and explanation/implication). The validity of the interpretive argument was strengthened to the extent which each type of evidence supports the inferences and assumptions regarding score interpretations and uses (Kane, 2006a, 2006b). A description of each stage is detailed in the next section.

## **Formative Stage: Formulating the Interpretive Argument and Assessment Development**

### **Developing the Interpretive Argument**

The initial interpretation of the test scores and uses of the proposed assessment was generated from the literature review considering the current need of the assessment. The following list of proposed claims was derived specifying what to measure and how to use the test results. This set of claims played an enabling role supporting an interpretive argument as types of evidence were investigated.

#### **Claims regarding the construct of IRS.**

1. The test measures students' level of IRS in two subdomains—ISI (informal statistical inference) and FSI (formal statistical inference).
2. The test measures IRS in the representative test domains.

#### **Claims regarding conclusions about the score interpretations and uses.**

3. The test produces scores with sufficient precision to be meaningfully reported.
4. The test is functional for the purposes of formative assessment.
5. The test provides information about students' level of IRS in the areas of ISI and FSI.

This set of claims laid out a sequence of the four inferences (scoring, generalization, extrapolation, and implication) leading from an observed test performance to the conclusions. The inferential network functioned as a framework encompassing all elements of the test design, development, and validation. Each of the four inferences is

described below, explaining what kinds of evidence were collected to support the inference.

***Scoring (the inference from observations of performance to an observed score)***. The degree of confidence about *scoring* inference provides information about the quality of the examinee's responses. As evidence, experts' judgments of the appropriateness of the answer key, testing conditions (e.g., how carefully and consistently the test was taken), and scoring methods in test specifications were gathered and analyzed. Item discrimination information from field-tests was also examined as a measure of score precision.

***Generalization (the inference from an observed score to the expected score on the universe of generalization)***. Validity of this inference can easily be evaluated in that the test was designed based on specified sub-domains with relatively homogenous items. The evidence supporting this inference included documentation of construct representation in a test blueprint, item discrimination, and item information function.

***Extrapolation (the inference from the universe score to the target score)***. This inference extrapolates from a narrowly defined universe of generalization to a score on a widely defined target domain beyond the test. The underlying assumption is that a score on the test reflects performance on a relevant target domain (students' actual level of IRS). Evidence supporting this inference included the test blueprint documenting content coverage, expert reviews, and think-aloud interviews. An examination of dimensionality also provided evidence to evaluate the validity of this inference, as it indicates whether the universe scores represent the unidimensional target score (IRS) or two-dimensional scores (ISI and FSI), as hypothesized from the literature.

***Explanation/implication (the inference from the estimated target score into a description of students' reasoning in statistical inference).*** This inference links the construct measured in the assessment to the description of the reasoning. This inference can be evaluated from a theory-based perspective since it needs evidence to show the extent to which the construct and performance (actual reasoning) are relevant to a specific discipline. An expert review on the test blueprint, item information and test information functions were examined.

The claims as assumptions and the list of inferences guided the set of comprehensive procedures in the test development and justification of score-based interpretations and uses. The procedures for the test development and validation are presented next.

### **Developing a Test Blueprint from the Literature Review (Theoretical Evidence 1: TE1)**

In a well-designed test blueprint, it is ensured that there is a sound relationship between the test contents in the blueprint and the construct the proposed test is intended to measure. Then, the test blueprint itself provides evidence based on the test content when it represents the content domain (AERA et al., 2002). In order to make an agreement on the test score interpretation and uses, it is required to decide on the scope of domains that will be covered in the assessment. However, since there is no criterion reference of IRS, the literature of informal and formal statistical inference was reviewed first. After the content domains were chosen, the types of reasoning to be assessed in the domains were specified based on what the previous researchers considered as important to be captured, which resulted in a preliminary test blueprint. Misunderstandings and

difficulties in statistical inference found in research literature were also categorized. The preliminary test blueprint is shown in Appendix B.

### **Expert Review of the Preliminary Test Blueprint (Theoretical Evidence 2: TE2)**

The preliminary test blueprint was reviewed by content experts, and evaluation reports were gathered to examine the adequacy of the test blueprint as a framework to represent the content domains. According to *Testing Standards*, qualified experts can judge the representativeness of the chosen test contents, and their judgments of the relationship between parts of the test and the construct also provide *evidence based on test content* (AERA et al., 2002). The experts who participated in the review process are described below, along with their credentials. The procedures of how they evaluated the preliminary blueprint follow.

**Participants.** The preliminary test blueprint developed from the literature was reviewed first by two internal experts, and then by three external experts. The internal experts are professionals in the program of statistics education at the University of Minnesota. To recruit external experts, the author contacted eleven potential professionals of statistics educators to ask them to evaluate the test blueprint in early May 2011. These reviewers were selected based on their background and research interests. It was also notable that the pool of reviewers has diversity in terms of their expertise and their level of teaching (*Testing Standards 1.7*, AERA et al., 2002). The email invitation letter and evaluation form were sent out to each of the potential reviewers, and three of them agreed to participate in the review process for both the test blueprint and assessment items. The consent form and invitation letter appear in Appendix C. All three reviewers were statistics educators who were actively engaged researchers in the area of statistics

education. The first reviewer has published many research studies about students' statistical inference, specifically utilizing technological tools or hands-on activities at the secondary and undergraduate levels in New Zealand.

The second reviewer's expertise is the development of statistics curricula, technological tools, and resources for teaching statistics. He has published in many research journals, specifically about how people elicit and acquire statistical reasoning at work. He is working in the Netherlands.

The third reviewer is an instructor in the Department of Statistics at a college in the Midwest area in the U.S. His expertise is in teaching rather than in research, but he has also been involved in several research projects about the topic of statistical inference. It was expected that his professional experience as a teacher of statistics would provide a valuable perspective in terms of a practical sense of assessing students' inferential reasoning. In addition, he was an introductory statistics textbook author who designed an innovative curriculum focused on developing IRS.

**Procedures.** During the entire process of developing a preliminary blueprint, the author had continuous discussions with the internal experts until an agreement was reached for the preliminary blueprint. Thus, only the reviews from the external experts are reported and analyzed in this paper.

Feedback on the preliminary test blueprint was collected from the three experts in late May 2011. Each reviewer was provided with a preliminary test blueprint and an evaluation form. The reviewers were asked to provide ratings for their agreement that the test blueprint was adequate as a framework to develop an instrument to assess the IRS in general (See the evaluation form for the questions in Appendix C.3). Specific evaluation



questions were also provided, asking the reviewers to rate the degree to which they agreed that the topics and learning goals documented in the blueprint represent the content domain (AERA et al., 2002). The reviewers were also asked to provide suggestions for changes if an item received a rating of less than 2. Items were judged to have a sufficient level of validity evidence if they had a mean rating of 3 (agree) or higher. For items with mean ratings of less than 3, the reviewers' suggestions for the item changes were carefully reviewed and discussed with an internal expert. In addition, the reviewers' comments on the free-response evaluation questions (e.g., whether there was anything missing from the content of the blueprint related to the constructs of informal and formal statistical inference) were also considered in revising the blueprint.

The feedback obtained from the reviewers was prioritized, restricting the topics and learning goals that would be included in the test blueprint. However, several times of individual meetings were held with the internal expert to discuss the reviewers' suggestions. To decide whether or not the suggested changes would be made in the blueprint, several aspects of the blueprint development were considered such as the score of the domains (statistical inference, ISI and FSI) delineated from the literature review and topics taught in introductory statistics courses in the U.S. As a result, the final version of the test blueprint was produced (See Appendix D).

### **Test Specifications (Theoretical Evidence 3: TE3)**

The author began test specifications by making a number of decisions regarding the test design. Most importantly, she attempted to develop measures of inferential reasoning in statistics, and not simply the contents described in textbooks. This corresponds to the reasoning and thinking explored mostly in the case study or qualitative

literature. Second, because no map of ISI or FSI existed in 2011, the researcher did not use any criterion reference of the instrument. Third, given projections regarding the size of the student sample to provide more accurate estimates of item properties, the author decided to use a multiple-choice (MC) format in the assessment items. While existing measurement tools such as observations, interviews, and discourse analyses would provide ample information in conceptualizing how students reason in given contexts, none were feasible for use in studies that would potentially include a large student sample. The items were designed to address a possible critique of the proposed assessment items—whether a multiple-choice format could measure cognitive and complex thinking, such as informal reasoning.

The critique has been noted by several researchers (e.g., Haertel, 2006) pointing out that the activities of reasoning and responding to a multiple-choice question are quite unlike the activities required in professional practice, such as an in-depth interview to probe an interviewee's reasoning (Haertel, 2006). Martinez (1999) also notes that scores on multiple-choice exams may reflect “test-wiseness”—an examinee's ability to recognize cues, to deploy response elimination strategies or to utilize other information in the stem to arrive at a correct answer without employing their actual reasoning of the underlying content being assessed. These potential threats to validity are of concern to the proposed assessment development. Similarly, constructed response (CR) items are considered to be more appropriate than MC items in assessing some cognitive thinking processes (e.g., mathematical reasoning studied by Traub & Fisher, 1977). Thus, it is appropriate to provide a rationale for use of the MC format in the proposed assessment.

**Rationale for MC format items in assessing cognitive thinking.** According to Haladyna (2004), the choice of an item format mainly depends on the kind of learning outcome it is intended to measure. In other words, in the process of measure specifications, we need to focus on the content and cognitive process. The proposed assessment included all MC items and the rationale for using the MC format is described below.

Validity arguments delineated from test scores and the score interpretation of the proposed assessment can be supported by examining the cognitive operations elicited by examinees (AERA et al., 2002). If the scores from MC and CR items provide the same degree of validity, either the MC format or the CR item could be used. If this is the case, then the MC format has advantages for several reasons: the MC format is more efficient in administration, objective scoring, automated scoring, and higher reliability (Haladyna, 2004).

An arguable issue is that CR and MC items elicit different mental behaviors: with higher levels of thinking, we feel comfortable using CR items because MC items are thought to elicit only lower levels of cognitive thinking. Martinez (1999) argues that this criticism has been aimed at the item writer, and not the test format. Haladyna (2004) also argues that with adequate training and practice, item writers can successfully write MC items with high cognitive demand. Hibbison (1991) provided empirical evidence that an MC test can capture higher levels of cognitive thinking, such as metacognitive, cognitive, and affective interactions. In terms of measuring the same construct with two different formats, Rodriguez (2003) provided a meta-analysis regarding the issue of the interpretability of test scores, either from the CR or MC format. He stated that MC and

CR item scores tend to be highly related when the content is intended to be similar. Therefore, it seems appropriate to use MC format in this assessment considering the intended uses of the proposed test.

#### **Developing an Item Pool (Theoretical Evidence 4: TE4)**

In order to develop the item pool, a set of items were examined from the six existing instruments—Statistical Reasoning Assessment (SRA, Garfield, 2003), the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS, delMas et al., 2006), Assessment Resource Tools for Improving Statistical Thinking Topic Scales on Test of Significance, Sampling Variance, and Confidence Intervals (ARTIST, Garfield et al., 2002), and R-PASS (Lane-Getaz, 2007). These instruments were selected since they include items assessing key types of conceptual understanding in statistical inference. After reviewing these instruments, some items assessing IRS were selected to be used as in the original resources or adapted from them.

In order to fill the gaps where no items exist in the category of the blueprint, the author reviewed assessments created for two curriculum projects (Beckman et al., 2010; Garfield, delMas, & Zieffler, in review), and a Test Bank for a textbook (Moore, Notz & Miller, 2008). For the items that are not a MC format, a CR format was used. This resulted in the preliminary assessment with a total of 36 items (31 MC items and 5 CR items).

#### **Expert Review for the Preliminary Assessment (Theoretical Evidence 5: TE5)**

The three experts who participated in the blueprint review process were asked to evaluate the preliminary assessment items. One of them was not available, so another expert was contacted. His expertise was in the area of statistics, and he had much

experience in teaching statistics at the college level. These three experts were asked to review the assessment by rating the extent to which each item reflected IRS, and whether an item measured ISI or FSI. The evaluation forms are presented in Appendix E. The three experts were asked to complete the general test evaluation form. The experts were also asked to report ratings of the extent to which they agreed that each item measures the specified learning goal. They were asked to suggest changes for items, if any. Suggested changes were reviewed carefully and discussed with an internal expert. Revisions based on expert review on the preliminary assessment resulted in the first version of the Assessment of Inferential Reasoning in Statistics (AIRS-1). This AIR-1 was used in the first cognitive interview of the summative stage described next section.

### **Summative Stage: Validating the Interpretative Argument**

In the summative stage, the validation process was focused on empirical checks of the inferences and assumptions in the interpretive argument. The validity of the interpretive argument was strengthened to the extent that the empirical checks support the inferences and assumptions made about score interpretations and uses. Different evidence sources were collected, and each of these is described below.

#### **First cognitive Interview Using Think-alouds (Empirical Evidence 1-1: EE1-1)**

Cognitive interviews were conducted at two different time points: before and after the pilot testing. These two interviews were conducted with two purposes: to change the CR items to MC items (the first cognitive interview), and to gather validity evidence based on response processes (the first and the second cognitive interviews). The first interview was conducted to capture variations of possible reasoning used by students to answer an item so that meaningful alternatives are constructed in a MC format. Students'

verbalized reasoning obtained from the first interview was also used to examine if students used correct reasoning when they chose correct answer to MC items (AERA et al., 2002).

It is necessary to verify that the assumed reasoning processes are actually elicited by test-takers, as opposed to contradictory processes (such as option elimination) that introduce construct-irrelevant variance into the scores (Messick, 1995). Ferrara et al. (2004) stated that mismatches between targeted and actual cognitive processing of test items undermine validity. Students' verbalization through cognitive interviews was used as evidence based on response process.

**Participants: First cognitive interview.** The participants were recruited in the middle of Summer 2011 from two sections of an introductory statistics course offered in Spring 2011. One of the two sections was taught online, and the other section was taught in a face-to-face environment. The researcher sent an email invitation letter to the students who had taken one of these two sections. The email invitation letter is presented in Appendix F.1. As an incentive for participation in a one-hour interview, a \$20 Amazon gift card was provided. Three female students out of 58 students agreed to participate in the interview. Two of the three students were from the face-to-face class section, and the other one was from the online section. The first two students were sophomores, and the third student was a senior. All of the three students were enrolled in liberal arts.

**Interview procedures.** As per the instructions contained in the email, the students called in for an appointment time. The author conducted the interviews by herself. At each of the think-aloud sessions, the author introduced herself and had the

student sign the consent form (see Appendix F.2). After the purpose and methods were briefly described, the author demonstrated the process of think-aloud.

The standardized think-aloud process was used to capture students' cognitive reasoning. Using the protocol developed by Ericsson and Simon (1993), two structured interview questions were asked before and after the student provided her/his reasoning—"What do you think this question is asking?" and, "Why do you think like that?" Minor interventions (e.g., "Can you talk about what you are thinking now?") were included to prompt students to think out loud if they appeared to be working on the problem, but not talking about it. The interview sessions were audiotaped.

At the beginning, a warm-up question ("How many windows do you think are in this building?") was first asked so that students could practice verbalizing their thinking processes when reaching an answer. The three students were interviewed with three different item sets of AIRS-1: the number of each item set ranged from 23 to 26 out of 35 in total, including 20 common items that all three students answered, and all of the 35 items were answered by at least one student. These three different item sets were provided in a counter-balanced presentation to control for test-taker fatigue (Schneider, Huff, Egan, Tully, and Ferrara, 2010). The common items asked of all three interviewees were—the CR items to be changed into the MC format; all of the items in the ISI part; and the items that require high cognitive demand in FSI. The 15 items presented only to one or two interviewees were either items asking for a simple understanding of a concept in inferential statistics or items that require a low cognitive demand in FSI—the items that are relatively obvious in terms of alignment between the response choice and

cognitive reasoning. Information obtained from think-aloud sessions was used to produce response choices in the MC format, resulting in the second version of the AIRS (AIRS-2).

**Framework used for analyzing the cognitive interviews.** To examine the degree of alignment between intended reasoning and actual reasoning elicited by students, a framework by Ferrara et al. (2003) was used. The researchers developed a framework to compare three types of item response demands: the intended item response requirements that the test designer-developers intend; the enacted item response requirements that the item writers build into the test items; and the actual cognitive processes that examinees actually use when they respond to the items. Within this framework, the author proceeded to the following three steps: determining the intended reasoning requirements; conducting a think-aloud with an expert to ensure that the intended reasoning requirements were enacted in test items; and collecting evidence regarding the examinees' actual cognitive thinking processes. An alignment between intended reasoning and students' elicited actual reasoning was then examined using a coding framework described in the next subsection.

An *intended reasoning* for each item was stated based on the learning goal developed in the item development stage. To verify that the intended reasoning was actually "enacted" in making a correct MC choice, one doctoral student in the statistics education program at the University of Minnesota was invited to perform a think-aloud from an expert view. She has been teaching introductory statistics for 2 years.

The expert's verbalized (*enacted*) reasoning and *intended reasoning* were first compared by examining whether the expert's reasoning process was aligned to the intended reasoning for each item. The analysis of the expert's reasoning from think-aloud



was conducted by the author using a holistic approach. *Actual reasoning* verbalized from the three think-aloud sessions was then compared to the *intended reasoning*.

**Coding framework.** The actual reasoning for each item verbalized by the three students was examined and compared to the intended reasoning for the item. The alignment between the intended and actual reasoning was coded to one of the four different categories: true positive (TP: correct answer choice and actual reasoning aligned with the intended reasoning); true negative (TN: incorrect answer and actual reasoning misaligned with the intended reasoning); false positive (FP: correct answer, but actual reasoning misaligned with the intended reasoning); and false negative (FN: incorrect answer, but actual reasoning aligned with the intended reasoning). These categories were slightly modified from an item demand analysis framework developed by Ferrara et al. (2004) and Schneider et al. (2010).

In the analysis of current study, TP indicates that the interviewee selected a correct MC response option and also the interviewee's actual reasoning aligned with the intended reasoning. TN indicates that the interviewee selected an incorrect MC response option, and also the interviewee's actual reasoning was misaligned with the intended reasoning. FP indicates that that the interviewee selected a correct MC choice, but the interviewee's actual reasoning was incorrect. Finally, FN indicates that the interviewee selected an incorrect MC choice, but the interviewee's actual reasoning matched the intended reasoning. Table 1 below simplifies this coding framework.

Table 1

*Think-aloud Coding Framework*

Actual reasoning Answer to MC	Matched to Intended reasoning	
	Yes	No
Correct choice	True Positive (TP)	False Positive (FP)
Incorrect choice	False Negative (FN)	True Negative (TN)

**A Pilot-test (Empirical Evidence 2: EE2)**

A pilot test was administered to one online section (N=23) of an undergraduate introductory statistics course in the Department of Educational Psychology at the University of Minnesota at the end of Summer 2011. The course was delivered online, and the assessment was administered as a final exam in an online test environment. A total of 2 hours was allowed to complete the assessment, and the time that each student took to complete the test was recorded in the online test system. Student response patterns were analyzed by examining the summary statistics of the correct-total scores. Item difficulties and item-total correlations were obtained as measures to examine the preliminary psychometric characteristics. Information drawn from the analysis of the pilot data was used for minor item revisions of the AIRS-2, and resulted in the AIRS-3.

**Second Cognitive Interview Using Think-alouds (Empirical Evidence 1-2: EE 1-2)**

Additional cognitive interviews were needed, since the CR items were changed to MC formats, and these items were not evaluated. In addition, the number of participants in the first cognitive interview was not representative of the general population, in that the students were recruited during the summer, when many students are out of town.

Moreover, the coding results for the alignment of the MC items revealed that most of these students showed a number of “True Negative” codes in alignment between their actual and intended reasoning, thereby indicating the interview participants’ lack of understanding of the concepts assessed by the test.

**Participants.** Two additional groups of students were contacted for the second cognitive interviews late in September 2011: one group from the same statistics class who participated in the pilot testing, and the other group from four sections of the same introductory statistics courses taught in fall 2011. For the first group, the author sent invitation letters to five students who got the five highest scores on the pilot test. Selecting the students with high scores was done to have a diverse group of interviewees in terms of ability level, given that the interviewees in the first interview setting did not provide good information for several items coded TN for all three previous interviewees.

One female student from the first group agreed to participate in the interview. This student’s score on the pilot test was within the highest 10%. Since the student had already taken the AIRS test during the pilot, a retrospective think-aloud was used (Ericsson & Simon, 1993). Five students from the second group participated in the second cognitive interview. These students were diverse in terms of their performance in the statistics course.

**Procedures and analysis.** The six students were asked with different item sets of AIRS-2: each interviewee answered between 23 and 26 (out of 34) items considering the limited time allowed and student fatigue. The procedures and the interview protocol for the think-aloud and the coding framework were similar to those conducted in the first interview (EE 1-1). The six item sets have 13 common items; these items were mostly the

items that showed TN coding in the first cognitive interviews in order to capture students' positive reasoning (either TP or FP). Each student's verbalized reasoning for each item was coded into one of the four categories (TP; TN; FP; FN).

**Inter-rater reliability analysis.** To examine inter-rater reliability, two other raters were invited to determine the accuracy (reliability) of the codes the author made. Both were doctoral students in the statistics education program at the University of Minnesota. One of the raters had been teaching introductory statistics courses for three semesters. The other rater has a master's degree in statistics and had taught an introductory statistics course for 2 years before coming to the statistics education program.

Since the nine students interviewed were asked different item sets, there were variations in the number of items interviewed. The range of the number of students interviewed for each item was between two and eight. For each item out of 34, two student-interviews conducted for that item were randomly selected without replacement, resulting in two interview sets, each consisting of 34 items. These two sets of interviews were randomly assigned to the two raters. The raters were trained to code items following the coding framework described above. After practicing with a couple of example items, the two raters completed the coding independently for their set of 34-items in one sitting. The codes the two raters made for each item set were then compared to the author's codes. For each set of interviews coded, two inter-rater agreement statistics were calculated: the percent of agreement between the two raters and Cohen's Kappa.

Results from the cognitive interviews and pilot testing informed further modifications on the items, mostly about wordings for clarification and formats. As a

result, a final version of the AIRS (AIRS-3) with 34 MC items was produced, and this was administered in a field test. Table 2 summarizes the major changes, the sources for changes made in each version of the assessment, and where the version was administered.

Table 2

*Changes in the AIRS Versions*

	Total number of items (# of MC items; # of CR items)	Changes made from:	Major changes implemented from the previous version	Administered for
Preliminary assessment	36 (31; 5)			Expert reviews
AIRS-1	35 (29; 6)	Expert reviews	3 MC items removed; 2 MC items added	An expert's interview and 1 <sup>st</sup> cognitive interviews
AIRS-2	34 (34; 0)	1 <sup>st</sup> cognitive interviews	All CR items changed to MC items	Pilot testing and 2 <sup>nd</sup> cognitive interviews
AIRS-3	34 (34; 0)	Pilot testing and 2 <sup>nd</sup> cognitive interviews	Wording changes for clarification	A large-scale administration

**Field-testing (Empirical Evidence 3: EE3)**

The 34 items of the AIRS-3 assessment were embedded in an online assessment tool that gave participants easy access to the test. A consent form and detailed instructions for the test were integrated into the online instrument (see Appendix G). Participants and detailed procedures of the online test are described below.

**Recruitment of instructors and test administration.** To recruit instructors to administer the online test, AIRS-3, the author sent invitation emails out to people who

were registered in one of three associations: AP statistics readers; the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) website (<http://www.causeweb.org>); and the Isolated Statisticians list serve (isoStat; <http://www.lawrence.edu/fast/jordanj/isostat.html>). A total of 46 statistics instructors agreed to administer the online test to their students for either part of the course grade or extra credit. A link to the online test was provided to the instructors with a unique code for their class. This unique code was given to identify the student scores for each instructor. One-hour was allowed for the students to complete the test, but the test was not timed. The online AIRS test was administered around the time of the final exam of each instructor's course in fall 2011.

**Participants: Students.** A total of 2,056 students took the AIRS test, and 1,978 students completed the test. These students were taught by 46 instructors in 39 higher education institutions across the United States. The majority of the students were enrolled at a university or a 4-year college, with about 17% of the students enrolled in 2-year colleges (see Table 3). Fifty-six percent of the students were female students, 37% were male students (nonresponse rate of 7%). Sixty-two percent of the students were Caucasian. Table 4 shows the pre-requisite mathematics courses for the statistics course in which students were enrolled. The largest group was represented by students enrolled in courses with a high school algebra requirement.

Table 3

*Types of Institution in a Large-Scale Assessment*

Institution Type	Number of Institutions (N=39)	Number of Instructors (N = 46)	Number of Students (%)	
2-year college	8	10	244	(12.3%)
4-year college	10	12	407	(20.6%)
University	21	24	1327	(67.1%)
Total	39	46	1978	(100.0%)

Table 4

*Mathematics Pre-requisites for the Statistics Course*

Mathematics prerequisite	Number of instructors (%)		Number of students (%)	
None	8	(17.4%)	553	(28.0%)
Algebra	17	(37.0%)	685	(34.6%)
College algebra	8	(17.4%)	334	(16.9%)
Pre-calculus	5	(10.9%)	157	(7.9%)
Others	3	(6.5%)	136	(6.9%)
Non-response	5	(10.9%)	113	(5.7%)

**Data analysis.** The response data obtained from field-testing were analyzed with respect to the different types of empirical evidence. First of all, since several items in the AIRS are in context-dependent item sets (24 items are in 8 contexts and 10 items are discrete), it is possible that items are not independent of each other. Thus, local item dependence (LID) was examined to determine an appropriate scoring method, as well as

to properly apply statistical techniques. Second, dimensionality in item responses was examined to determine if responses revealed the hypothesized structure of the assessment (two-factor structure with ISI and FSI), as developed from the literature and verified from expert reviews. Third, the item responses were fitted to an appropriate IRT model selected from the results of the previous two analyses—examination of LID and dimensionality. Each of these analyses is detailed below.

*Local item dependence (LID).* The AIRS test has some context-based items that include a component of variation that is attributable to the contexts. That component of variation induces local dependence among the items that follow each context. Local item dependence (LID) occurs when respondents' answers to a particular item depend not only on their standing on the latent trait, but also on their responses to other items (de Ayala, 2009). There are several potential reasons that LID arises: sharing a common passage, content, knowledge, item chaining, speediness, fatigue, practice effects, and item or response format (Yen, 1993); the physical layout of the test booklet (Muraki & Lee, 2001). An examination of LID is necessary before conducting other statistical analyses, since the presence of LID may result in an inaccurate estimation of item parameters, test statistics and examinee proficiency (Fennessy; 1995; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989), thus introducing an additional (unintended) dimension into the test (Wainer & Thissen, 1996), and overestimating reliability estimates and test information functions (Thissen et al., 1989; Sireci et al., 1991). When seemingly distinct items related to a context exhibit dependency, grouping them together into a testlet more properly models the test structure. Using this strategy, local item independence holds across testlets, since the testlet is modeled as a unit (i.e., a



polytomous item). Moreover, considering that the responses will be scored using IRT analysis, fitting sets of locally dependent items as testlets models the testlet-based structure of the test in a way that meets the local independence assumption of IRT.

Among several different methods for assessing LID in dichotomous data, two methods were employed: reliability analysis and Local Dependence indices for item pairs (LD indices, Chen & Thissen, 1997). The two methods are described in detail below.

*Reliability analysis to detect LID.* In a reliability analysis, context-dependent item sets were modeled using testlets so that each testlet includes all of the dependent items in terms of the context. However, the first testlet (TL1) is divided to two different testlets (TL1-1 having items 3 to 6 and TL 1-2 having items 7 and 8) based on the learning outcomes that the questions of each testlet measure. This was also done to address the possible loss of information when one “large” testlet is created, since many “small” testlets are likely to retain more information than one “large” testlet (Yen, 1993). In the reliability analysis, two coefficient-alphas were compared between the one when the test is considered to only comprise locally independent (dichotomous) items, and the other one for testlet-based items (Green, Bock, Humphreys, Linn, & Reckase, 1984; Zenisky, Hambleton, & Sireci, 1999). For dichotomous data, traditional scoring was used by treating all items as discrete, and thus, independent. For the testlet polytomous data, an examinee’s score on a testlet was computed by adding up the number of items within the testlet that the person answered correctly. Table 5 summarizes the structure of the testlet-based test format in terms of the number of items.

Table 5

*Structure of the Testlet-Based Test*

Testlets (TL)	Item Number	Number of Items
TL1-1	Item 3-6	4
TL1-2	Item 7 – Item 8	2
TL2	Item 9 – Item 11	3
TL3	Item 12 – Item 13	2
TL4	Item 15 – Item 16	2
TL5	Item 19 – Item 20	2
TL6	Item 21 – Item 22	2
TL7	Item 24 – Item 26	3
TL8	Item 27 – Item 30	4
Ten discrete items	Item 1, 2, 15, 17, 18, 23, 31, 32, 33, 34	10
Total items	9 testlets and 10 discrete items	34

In comparing the coefficient-alpha, a lower reliability coefficient for the testlet data compared to the one for the dichotomous data might indicate an overestimate of the latter coefficient (Sireci et al., 1991, Thissen et al., 1989). However, lower reliability of testlet data could be due to the fact that the number of items in the testlet data is less than those in dichotomous data. Therefore, the Spearman-Brown formula was employed as a way to compare the reliability of discrete 34 item responses and the reliability of testlet responses with respect to the effect of test length on the reliability (Sireci et al., 1991; Wainer, 1995). This statistic is commonly used to predict the reliability of a test after changing the test length. This relationship is particularly useful in examining the presence

of LID in that it allows us to determine whether the overestimate of the reliability in discrete response data is due to the presence of LID or due to the greater number of items. It also provides information of which scoring method is more reliable and useful. The ltm package and the CTT package in R were used to obtain reliability estimates and the spearman-brown coefficients.

*Likelihood Ratio G2 statistic: Local dependence indices for item pairs.* The reliability analysis to detect LID described above is useful to examine the presence of LID in the item responses as a whole. However, this method does not provide information about which pairs of items are dependent, a necessary step to confirm that the items within the same passage show high correlations, and also to determine which items need to be clustered as a testlet in scoring. Chen and Thissen (1997) proposed the LD index, which provides a straightforward analysis of pair-wise measures of association between responses to item pairs. These pair-wise measures have been found to be more powerful than test- and item-level measures in detecting misfits for unidimensional IRT models. The LD indices are based on 2X2 contingency tables. For each pair of dichotomous items  $i$  and  $j$ , the following two contingency tables can be constructed. In Table 6,  $O_{pq}$  is the observed frequency and  $E_{pq}$  is the expected frequency, where 1 and 0 present the correct and incorrect responses, respectively, and  $E_{pq}$  is predicted by the IRT model.

Table 6

*Contingency Tables of Observed- and Expected Frequency*

		Item j	
		0	1
Item i	0	O <sub>11</sub>	O <sub>12</sub>
	1	O <sub>21</sub>	O <sub>22</sub>

		Item j	
		0	1
Item i	0	E <sub>11</sub>	E <sub>12</sub>
	1	E <sub>21</sub>	E <sub>22</sub>

A Pearson's  $\chi^2$  index is then computed as:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

And, the corresponding likelihood ratio  $G^2$  statistic is computed as:

$$G^2 = -2 \sum_{i=1}^2 \sum_{j=1}^2 O_{ij} * \ln \frac{O_{ij}}{E_{ij}}.$$

These two LD indices are distributed as  $\chi^2$  with degrees of freedom of 1 when the assumption of local dependence is held. Chen and Thissen (1997) found the observed 95<sup>th</sup> percentiles of the  $\chi^2$  and  $G^2$ -LD indices under the null condition (local independence), and suggested 3.84 as a critical value to flag item pairs as locally dependent if the obtained index exceeds 3.84.

This method is particularly useful in this study in that the results of LD indices would provide information of how items should be combined into testlets. This would then allow the same IRT technique to be used for test scoring. Chen and Thissen's LD indices for 34 dichotomous response data were computed fitting the full-information unidimensional factor model. The mirt package in R was used to fit the IRT model and to obtain LD indices.

*Dimensionality.* Confirmatory Factor Analysis (CFA) was conducted to assess the dimensionality of the AIRS response data. The AIRS items were decomposed after specifying a theoretical structure in terms of ISI and FSI. Mplus (Muthén & Muthén, 2010) was used for the CFA. For the two data sets (discrete response data and testlet-based data), two factor models (a unidimensional model and a two-factor model) were examined and compared in terms of standardized regression weights (factor loadings) and fit indices.

*IRT analysis.* The results obtained from the analysis of LID indicated that testlet-based polytomous data were more appropriate due to the presence of LID in the original dichotomous data. Thus, analyses of item parameters and item information were performed with the polytomously scored testlet-based responses. Item response theory was employed in the analysis, and it is detailed below.

*Item response theory.* In examining item quality, as well as test performance, item response theory (IRT) is considered as the standard, if not preferred method of conducting psychometric evaluations of new and established measures (Embretson & Reise, 2000; Fries, Bruce, & Cella, 2005; Lord, 1980). Among the many advantages of IRT over classical test theory (CTT), IRT addresses three problems inherent in CTT.

First, IRT overcomes the problem of item-person confounding found in CTT. IRT analysis yields estimates of item difficulties and person-abilities that are independent of each other, whereas in CTT, item difficulty is assessed as a function of the abilities of the sample, and the abilities of respondents are assessed as a function of the item difficulty (Bond & Fox, 2001). Second, the use of categorical data may violate the scale and distributional assumptions of CFA (Wirth & Edwards, 2007), which may result in biased model parameters. Third, the IRT approach to the standard error of measurement has several benefits: (a) the precision of measurement can be evaluated at any level of the latent trait instead of averaged over trait levels as in CTT; and (b) the contribution of each item to the overall precision of the measure can be assessed and used in item selection (Hambleton & Swaminathan, 1985).

*Model identification and estimation method.* The GRM is a type of polytomous IRT model, an extension of Thurstone's (1928) method of successive intervals to the analysis of graded responses on educational tests. This model was first discussed by Samejima (1969).

The GRM specifies the probability of a person responding with a category score  $x_j$  or higher versus responding in lower category scores. In other words, the GRM specifies the probability of a person responding in category  $k$  or higher versus responding in categories lower than  $k$ . Responses to item  $j$  are categorized into  $m_j+1$  categories, where higher categories indicate more of the latent trait. According to the GRM, the probability of obtaining  $x_j$  or higher is given by

$$P_{x_j}^*(\theta) = \frac{e^{\alpha_j(\theta - \delta_{x_j})}}{1 + e^{\alpha_j(\theta - \delta_{x_j})}}.$$

where  $\theta$  is the latent trait,  $\alpha_j$  is the discrimination parameter for item  $j$ ,  $\delta_{x_j}$  is the category boundary location for the category score  $x_j$ , and  $x_j = \{0, 1, \dots, m_j\}$ .

In the GRM, score categories are separated by category boundaries: for the case that testlet-based responses have five categories (resulting from combining 4 discrete items), the five score levels are separated by four category boundaries: the boundary between score level 1 and 2, 2 and 3, 3 and 4, 4 and 5, respectively. Under the GRM model, each item has a discrimination parameter and a set of  $m$  threshold parameters when there are  $m+1$  categories. Therefore, the subsequent threshold parameters distinguish the probabilities of scoring less than score category  $k$  and greater than or equal to score category  $k$ .

For the testlet-based response data in this study, the GRM was employed for hierarchically ordered response categories by allowing the discrimination parameter to vary across items (or testlets) and between response categories. The `ltm` and `irttoys` packages in R were used to estimate item parameters. Package `ltm` fits the GRM under the logit link using Marginal Maximum Likelihood Estimation (MMLE). The estimates of item parameters (discrimination, category threshold), item (testlet) information, and item (category) characteristic curves were provided. Table 7 summarizes the study phases and timeline.

Table 7

*Summary of Data Collection Phases*

Phase	Sources of Validity Evidence	Participants	Product and Analysis	Time Line
Formative Stage: Theory-based evidence	TE1. Literature review to develop a test blueprint	Author	Preliminary test blueprint	Fall 2010
	TE2. Expert review on the preliminary test blueprint	Three reviewers	Final version test blueprint	March, 2011
	TE3. Test blueprint and test specifications	Author	Item specifications	
	TE4. Literature review to develop an item pool	- Existing instruments assessing statistical reasoning - Test blueprint	Preliminary assessment (36 items: 31 MC type and 5 CR type)	
	TE5. Expert reviews on the preliminary assessment	Three content experts	1 <sup>st</sup> version assessment (AIRS-1: 35 items, 29 MC type and 6 CR type)	April, 2011
Summative Stage: Empirical evidence	EE1-1. First cognitive interview	One expert's cognitive interview; Undergraduates in introductory statistics courses at the U of M (N=3); Sample 1	2 <sup>nd</sup> version assessment (AIRS-2: 34 MC items)	Early May, 2011
	EE2. A pilot test	Students who have taken an introductory statistics course at the U of M (N=23)	3 <sup>rd</sup> version assessment (AIRS-3: 34 MC items)	Summer, 2011
	EE1-2. Second cognitive interview	Sample 1 in EE1-1 (N=3); Undergraduates in intro stat courses at the U of M (N=6)	Alignment between intended reasoning, enacted reasoning, and actual reasoning.	Summer and Fall, 2011 (cont.)



Phase	Sources of Validity Evidence	Participants	Product and Analysis	Time Line
<i>Table 7, cont.</i>				
EE3.Field test	Undergraduates who are taking statistics courses in U.S institutions (N = 1,978 )	Factor analysis, Examination of local independence, IRT analysis	Fall, 2011	

## Chapter 4

### Results

This chapter discusses the results of the study. The data analysis is described along with the structure used in data collection procedure—a formative stage and a summative stage. The first section presents the analysis results of theoretical evidence obtained from formative stage. Developmental process of test blueprint and assessment is also presented in terms of the changes made in the previous version of the instrument. Results of the analysis for empirical evidence gathered in summative stages are examined. This chapter ends up with synthesis of the study results integrating all of the theoretical and empirical evidence sources. Underlying inferences about test uses and score interpretations are evaluated by judging the claims laid out in the formative stage. The four inferences (scoring, generalization, extrapolation, and explanation) are revisited and examined by evaluating the plausibility of the claims.

#### **Analysis Results for the Data Obtained in the Formative Stage**

##### **Results from the Literature Review to Create the Test Blueprint: Theoretical Evidence 1 (TE1)**

**A test blueprint developed from the literature review.** The initial test blueprint was built from the literature about IRS. Representing the content domains of IRS, the literature was centered around two areas: Informal statistical inference (ISI) and Formal statistical inference (FSI). These two content areas were used as hypothetical structure of a construct IRS providing the scope of the content to be covered in the assessment.

The definitions of the construct IRS, and two content domains ISI and FSI, which have been clarified in the previous chapter, were revisited. In this study, ISI was defined

as a domain of statistical inference that involves informal processes of making arguments to support inferences about unknown populations based on observed samples not necessarily using standard statistical procedures. FSI was defined as a domain of statistical inference that involves making a conclusion about population from samples or to formally test hypotheses, using standard statistical methods. As reviewed in Chapter 2, the topic category of sampling distribution was considered to represent foundations of statistical inference. The topic of hypothesis testing was used as the second category representing the concepts and ideas of formal statistical inference. Therefore, two content areas of FSI were considered as the main topics in this domain—sampling distributions and hypothesis testing. As a result, the domains of the blueprint were categorized into three areas: informal inference (Inf), sampling distribution (SD), and hypothesis testing (HT).

For the topic of sampling distributions, five content domains were culled from the literature: the concepts of samples and sampling; the Law of Large Numbers; population distribution and frequency distribution; population distribution and sampling distribution; and the Central Limit Theorem. The literature review resulted in a preliminary test blueprint. Table 8 presents some examples of the content domains, topics, and learning goals of ISI and FSI. The preliminary test blueprint is shown in Appendix B.

Table 8

*Examples of the Preliminary Blueprint*

Test blueprint to assess informal inference				
Category	Content Domains	Learning Goals		Literature
Informal Inference (Inf-1)	Uncertainty	Being able to express uncertainty in making inference using probabilistic (not deterministic) language		Makar and Rubin (2009), Zieffler et al. (2008)
Inf-2	Aggregates	Being able to able to reason about a collection of data from individual cases as an aggregate		Makar and Rubin (2009); Rubin, Hammerman, & Konold (2006); Pfannkuch (1999)
Test blueprint to assess formal inference				
Category	Content Domains	Learning Goals	Misconceptions <sup>a</sup>	Literature
Sampling distribution (SD-1)	Samples and sampling	-Understanding the definition of a sampling distribution -Understanding the role of sampling distributions	A tendency to predict sample outcomes based on causal analyses instead of statistical patterns in a collection of sample outcomes	Saldanha and Thompson (2002); Saldanha (2004); Rubin, Bruce, and Tenney (1991)
SD-2	Law of Large Numbers (Sample representativeness)	Understanding that the larger the sample, the closer the distribution of the sample is expected to be to the population distribution	A tendency to assume that a sample represents the population, regardless of sample size ( <i>representativeness heuristic</i> )	Kahneman and Tversky; Rubin et al. (1991); Saldanha & Thompson (2002); Metz (1999); Watson & Moritz, (2000a, 2000b) ( <i>cont.</i> )

Category	Content domains	Learning goals	Misconceptions <sup>a</sup>	Literature
<i>Table 8, cont.</i>				
Hypothesis testing (HT-1)	Hypothesis testing	-Being able to describe the null hypothesis -Understanding the logic of a significance test	-Failing to reject the null is equivalent to demonstrating it to be true (Lack of understanding the conditional logic of significance tests) -Lack of understanding the role of hypothesis testing as a tool for making a decision	Batanero (2000); Nickerson (2000); Haller & Krauss (2002); Liu & Thompson (2009); Vallecillos (2002); Williams (1999); Mittag & Thompson, 2000
HT-2	<i>P</i> -value and statistical significance	Being able to recognize a correct interpretation of a <i>P</i> -value	Misconception: <i>P</i> -value is the probability that the null hypothesis is true and that (1- <i>p</i> ) is the probability that the alternative hypothesis is true	Carver (1978); Falk & Greenbaum (1995); Nickerson (2000)

<sup>a</sup>*Note.* Misconceptions of the topic of ISI have not been found in the literature since empirical research on the topic of informal statistical inference has not been investigated.

## **Expert Review of the Preliminary Test Blueprint: Theoretical Evidence 2 (EE2)**

**Results of evaluation ratings.** Three professionals in statistics education provided their feedback and suggestions on the preliminary test blueprint. Table 9 presents the results of the experts' ratings for each evaluation question.

As shown in the table in the next page, the experts generally agreed that the content domains and learning goals listed in the preliminary blueprint represent the target domains of ISI and FSI. It also appeared that the learning goals identified are adequate to assess students' ISI and FSI. However, there are two evaluation questions that one expert assigned to "*disagree*": question 4 and question 8. The expert provided comments for these ratings, and these are detailed below along with the general and specific comments.

Table 9

*Results of Expert Review on Test Blueprint*

Item	Evaluation Questions	Ratings Made by Experts			
		Strongly Agree	Agree	Disagree	Strongly Disagree
1	The topics of the blueprint represent the constructs of <i>informal</i> statistical inference.	X	XX		
2	The topics of the blueprint represent the constructs of <i>formal</i> statistical inference	X	XX		
3	The learning goals of the blueprint are adequate for developing items to assess students' understanding of <i>informal</i> statistical inference.	X	XX		
4	The learning goals of the blueprint are adequate for developing items to assess students' understanding of <i>formal</i> statistical inference.	X	X	X	
5	The set of learning goals is well supported by the literature.	X	XX		
6	The learning goals are clearly described.		XXX		
7	The categories of the blueprint are well structured.		XXX		
8	The blueprint provides a framework of developing a test to assess informal and formal statistical inference.	X	X	X	

**Results of the suggestions and comments.** In addition to the ratings for the validity questions to evaluate the test blueprint, the experts were also requested to identify any important content domains in ISI and FSI not listed in the blueprint. It was asked to comment about any redundancy, and to provide additional suggestions to improve the test blueprint.

There were common suggestions made from two reviewers. First of all, reviewers 1 and 2 suggested including real world applications in the blueprint. Reviewer 1 commented, “There is no attention to the inferences about the real world or contextual knowledge” in the current version. It was also suggested that the current blueprint had too much focus on the “limited population” in the categories of SD (sampling distribution) and HT (hypothesis testing; Reviewers 1 and 3). One of the reviewers noted, “One can conceptualize a process as an infinite, undefined population.” Similarly, another reviewer commented that there is no content from an experimental perspective saying, “It only talks about samples from limited populations.” Another common suggestion was provided about the topic of “effect size” (Reviewers 2 and 3). In the category of HT-2, the topic covers definitions of *P*-value and statistical significance. In addition to the *P*-value, a reviewer suggested to include consideration of “how large is the effect,” which is related to the concept of the effect size. A similar comment was made by another reviewer with a suggestion of adding the “data quality or soundness of the method” to the current blueprint.

Specific suggestions were also provided regarding additional topics to be included in the test blueprint. The topics are:

- Correlation and regression (Reviewer 1)
- Using models in ISI (Reviewer 1)
- Using meta-cognitive awareness of what inference is as opposed to performing procedures (Reviewer 1)
- Confidence intervals (Reviewer 2)



- In the category of HT-6, add designing a test to compare two groups in an experiment, not just from populations (Reviewer 2)
- Consider including randomization and bootstrapping methods (Reviewer 2)
- In the category SD-2, include “biased sampling” for sampling representativeness (Reviewer 3)

These suggestions were reviewed carefully by the author, and were also reviewed with an internal advisor. Discussion between the author and internal advisor centered around whether or not these topics should be included. The definition and the domains that the proposed assessment targets were prioritized for the decision. Table 10 summarizes the changes implemented from the reviewers’ comments. The rationale for whether those comments were implemented or not appears in Appendix H.

Table 10

## Changes to Test Blueprint Implemented from Expert Reviews

Category	Changes Suggested	Changes Made in the Blueprint
Inf	Include real world or contextual knowledge	Added some learning goals to <i>inferential reasoning in a given context</i>
Inf	Include learning goals about “Using models in informal inferential reasoning”	In two categories, informal inference and formal inference, the learning goals of setting up the null model in a given context was added
Inf	Include using meta-cognitive awareness of what inference is as opposed to performing some techniques	Not included in the blueprint
SD and HT	Too focused on the limited population: Add a process as an infinite (undefined) population; Add statistical testing in experiments	Added the topic categories, DE (designs of study) and EV (evaluation of study) to capture students’ understanding of the characteristics of different types of studies
HT	Include the learning goals about an understanding of effect size	In a new category of EV, added the learning goal, “Being able to evaluate the results of hypothesis testing considering —sample size, practical significance, effect size, data quality, soundness of the method, etc.”
HT	Include data quality, soundness of the method etc.	The topic category, “Evaluation of HT (EV),” was separated out from the Hypothesis Testing categories since this topic is more about assessing how to interpret and evaluate the results from statistical testing by integrating different kinds of information in a given study (e.g., random assignment, sample size, data quality). The learning goal about, “Being able to evaluate the results of hypothesis testing (considering sample size, practical significance, effect size, data quality, soundness of the method, etc.),” was included in this EV category.

(cont.)

Category	Changes Suggested	Changes Made in the Blueprint
<i>Table 10, cont.</i>		
SD or HT	Include a topic category on Confidence Intervals	The topic category, “Inference about Confidence Interval, CI” was added.
SD -2	Add a topic of recognizing “biased sampling” for sampling representativeness	The topic of the “Law of Large Numbers” was changed to “sample representativeness” to assess whether students realize the importance of unbiased sampling (quality of samples), in addition to a large sample (sample size)
HT-6	Add designing a test to compare two groups in an experiment	In ST-3 (changed from a category of HT), the learning goal, “designing a statistical test to compare two groups in an experiment,” was added.
HT	Include randomization and bootstrapping methods	Not included as a separate learning goal, but will be assessed in a way so that items get at students’ reasoning about the ideas involved in randomization and bootstrap methods.  Considering that hypothesis testing based on a normal distribution-based approach is not the only way of statistical testing, the original category about hypothesis testing (HT) was changed to statistical testing (ST), which includes randomization or bootstrap methods.
In general	Add the topics, correlation and regression	Not included in the blueprint since the suggested topics were considered as not being in IRS defined in this study.

There were topics that the reviewers suggested to include that were not implemented in the blueprint. For example, one reviewer suggested adding content about “correlation and regression.” However, these were considered as *literacy* or part of descriptive statistics rather than a topic of inferential reasoning. Another reviewer commented that ISI might also include “meta-cognitive awareness”, but we decided that the topic of meta-cognition does not fit the definition of ISI. In addition, there was no literature found regarding this topic as part of ISI. The changes made from the expert reviews resulted in the final version of the blueprint (See Appendix D). In the last review process of the blueprint, the acronyms representing the topic categories, SD (sampling distribution) and HT (hypothesis tests), were changed to SampD and Stest, respectively, to avoid confusion: in statistics, the acronym of SD is mostly used to represent standard deviation. The final version of the blueprint was used to develop the preliminary version of the assessment.

### **Test Specifications: Theoretical Evidence 3 (TE3)**

In the *Testing Standards*, it is recommended that test specifications are detailed before the test development, and items are developed along with the test specifications (AERA et al., 2002). Decisions on the specifications were made primarily from the previous steps—literature review, test blueprint, expert reviews on the blueprint, and final review and discussion with an internal expert. The following list presents the test specifications made from the previous steps. From the review of literature and experts, it was decided that the content domains of IRS include the content categories of—sampling distribution (SampD), statistical testing (Stest), confidence interval (CI), and evaluation of the study (EV). Considering the scope of the content coverage, item format, and

feasibility of the test administration, 30 to 35 items were proposed as an appropriate number of test items. As the item format of the final version assessment, a MC format is used given the topic coverage, the desired sample size to be collected, and efficiency and accuracy of scoring. It was also considered that item responses obtained from a MC format item can be analyzed using modern psychometric theory providing ample information about item quality as well as test information. As appropriate amount of time for taking the test, 60–90 minutes will be given to students considering the feasibility of the test administration for instructors, desired difficulty, and student fatigue. The test will be administered online, with instructions presented on the front page. Individual scores will be scored automatically and these scores will be reported as a correct-total score.

#### **Examining Existing Instruments and Literature for Developing Preliminary Test:**

##### **Theoretical Evidence 4 (TE4)**

From existing instruments (SRA, ARTIST topic scales, CAOS, and RPASS), 10 items were selected that matched the learning goals in the blueprint. Two items were selected from the Sampling Variability topic scale from the ARTIST website, and 8 items were selected from the CAOS test. Although there are some items asking about statistical inference in the other instruments—SRA, RPASS, and the other topic scales from ARTIST (Confidence Interval topic scale, Test of Significance topic scale)—these items were judged to not be assessing inferential reasoning.

Of the 10 items adopted from existing instruments, 5 items were used as in the original instruments. For the other 5 items, 2 items modified by Ziegler (2012) were used. The other 3 items were revised by the author and Robert delMas adopting the contexts from CAOS. These 10 items were matched to the 13 learning goals (out of 38 learning

goals total) listed in 9 topic categories (out of 18 topic categories). Details for the changes made from the original items and the rationale for the changes are appeared in Appendix N.

The gaps shown in the blueprint (25 learning goals in 9 topic categories) were filled from reviews on two research projects and a test bank of a textbook. Nine items were made from revisions of interview questions used in the CATALST project (Garfield et al., in review). Six items were adopted from the assessment developed for a curriculum evaluation at UCLA (Beckman et al., 2010). Ten items were adapted from the test bank written by textbook authors (Moore et al., 2008). One item was created by the author from a discussion with Robert delMas. The original resources for the preliminary test are summarized in Table 11.

Table 11

*Resources of Items in a Preliminary Version*

Type of Resource		Item Numbers (in preliminary assessment, Appendix I.1)	Original Resources	Number of Items
Existing instruments	ARTIST	13	Adapted items from ARTIST Sampling variability topic scale	1
	CAOS	2, 14, 15, 36	Adapted items from CAOS 7,17, 34, 35	9
		16, 22	Adapted contexts from CAOS 32 and 37 and 2 items created by the author and an advisor	
		10	Adapted and merged from three items in CAOS 11-13	
	18-19	Adapted from a research study by Ziegler's research project as adapted from CAOS 23, 24		
Other resources	Research project or a textbook	1	Adapted from Konold & Garfield (1993) as adapted from Falk 1993 (problem 5.1.1, p. 111)	26
		3-9, 11-12	Adapted and revised from CATALST project	
		20-21	Adapted from UCLA Evaluation project (Robert Gould)	
		23-25	Adapted from CSI project (Rossman & Chance) as adapted for use in Robert Gould Evaluation project (Beckman et al.)	
		17, 26-29, 30-31, 33, 34, 35	Adapted from Instructor's Manual and Test Bank for Moore and Notz' (Moore et al., 2008)	
		32	Created by the author and an Robert delMas	Total 36 items

**Expert Review for the Assessment Items: Theoretical Evidence 5 (TE 5)**

The three expert reviews on the preliminary version of the assessment were examined. Data from experts’ reports on two item evaluation forms were analyzed: one for general evaluation of the test, and the other for evaluation of each item in the test. Table 12 presents a summary of evaluations that three reviewers reported for the test. For item evaluation, two questions were asked for each item: 1) the extent to which the specified learning goal that the item assesses is related to informal (or formal) statistical reasoning; and 2) the extent to which the item is appropriate to assess the targeted learning goal. Table 12 shows the items that at least one expert rated either “Strongly Disagree” or “Disagree”.

Table 12

*Items rated "Strongly Disagree" or "Disagree" by at least One Reviewer*

Learning Goals	Please check the extent to which you agree or disagree with each of the following statements.	Items that at least one expert rated either “Strongly Disagree” or “Disagree”
Evaluation question	This learning goal that this item gets at is related to informal (or formal) statistical reasoning.	Item 5, 7, 12, 13, 20, 21, 28, 33
	This item is appropriate to assess the learning goal aimed.	Item 7, 9, 12, 21, 28

In addition to the quantitative ratings to the Likert-scale evaluation questions, changes were suggested for the items rated either as “strongly disagree” or “disagree.” Table 13 presents the original item, the reviewer’s comment, and the changes made for



the item, for the items that had at least one rating of “disagree” or “strongly disagree”.  
(See Appendix J for detailed description of the reviewers’ suggestions and comments).

Table 13

*Changes made for the Items Rated "Strongly Disagreed" or Disagreed"*

---

[Original item 5] A statistician wants to set up a probability model to examine how often the result of 5 B's out of 10 spins could happen with the spinner just by chance alone. What would be the probability model the statistician can use to do a test? Please describe the null model.

- a. The probability for each letter is  $p(A)=1/4$ ,  $p(B)= 1/4$ ,  $p(C)=1/4$ ,  $p(D)=1/4$ .
- b. The probability for letter B is  $1/2$  and the other three letters each have probability of  $1/6$ .
- c. The probability for letter B is  $1/2$  and the probabilities for the other letters sum to  $1/2$ .

[Experts' comment on item 5]

Expert 1: The distracters seem to be very implausible. Might need to have pilot testing using a free-response format.

Expert 2: Add this: "trials are independent of each other."

[Changes made for item 5 ] This item was changed to a CR format to recreate plausible alternatives.

*(cont.)*

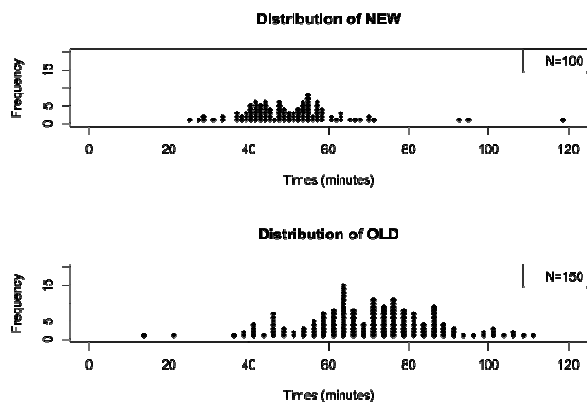
---

---

Table 13, cont.

[Original item 10]

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



- a. The old formula works better. One person who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.
- b. The average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief on average about 20 minutes sooner than those taking the old formula.
- c. We can't conclude anything from these data. The number of patients in the two groups is not the same, so there is no fair way to compare the two formulas.

[Expert's comment on item 10] The CAOS test has these as three separate items, and students indicate if they think each statement is Valid or invalid. You get more information about the students' thinking if you have them respond to the validity of each statement. You could also then see if a single score based on their responses to all three items provides more information than a separate score for each item.

[Changes made for item 10] This item was separated as three MC items; two items were added.

(cont.)

---

---

Table 13, cont.

Original item 13] A random sample for different courses taught at a University is obtained, and the mean textbook price is computed for the sample. To determine the probability of finding another random sample with a mean more extreme than the one obtained from this random sample, you would need to refer to:

- a. the distribution of textbook prices for all courses at the University.
- b. the distribution of textbook prices for this sample of University textbooks.
- C. the distribution of mean textbook prices for all samples from the University.

[Expert's comment] You need to add "of size 25" to this part.

[Change made for item 13] In option C, the distribution of mean textbook prices for all samples of size 25 from the University.

[Context of original items 20 -21] **Read the following information to answer questions 20 and 21:**

Data are collected from a research study that compares performance for professionals who have participated in a new training program with the performance for professionals who haven't participated in the program. The professionals are randomly assigned to one of two groups, with one group being given the new training program, and the other group being not given. For each of the following pairs of graphs, indicate what you would do next to determine if there is a statistically significant difference between the training and no training groups.

[Expert's comment] You need to give the sample sizes for both groups and state what the time is measuring.

[Change made for items 20-21] ... The professionals are randomly assigned to one of the two groups, with one group receiving the new training program (N=50) and the other group not receiving the training (N=50).

[Original item 28] The report of the study states, "With 95% confidence, we can say that the average score for students who take the college admissions test a second time is between 28 and 57 points higher than the average score for the first time." By "95% confidence" we mean:

- a. 95% of all students will increase their score by between 28 and 57 points for a second test.
- b. We are certain that the average increase is between 28 and 57 points.
- c. We got the 28 to 57 point higher mean scores in a second test in 95% of all samples.
- d. 95% of all adults would believe the statement.

[Expert's comment] Option C should be reworded to better capture ideas about population differences.

[Change made for item 28]

- c. 95% of all students who take the college admissions test would believe the statement.
  - d. We are 95% certain that the average increase in college admissions scores is between 28 and 57 points.
-

The suggested changes were reviewed and implemented resulting in the first version of the assessment, titled Assessment of Inferential Reasoning in Statistics (AIRS-1). This version consisted of 35 items (29 MC items and 6 CR items). AIRS-1 was used in the first cognitive interview of the summative stage.

### **Analysis of Results in the Summative Stage**

Evidence gathered in the summative stage was used for empirical checks of the inferences and assumptions in the interpretive argument structured in the formative stage. The cognitive interview results from an expert are first described in terms of whether or not the expert's elicited reasoning matched the intended reasoning for each item. Cognitive interviews with students were conducted at two different time points with two different purposes, respectively: to change CR items to MC items based on student response variations, and to collect validity evidence based on response processes. The 34 MC items were piloted to gather preliminary information about item quality, appropriateness of test specifications, and response patterns. Results from the test pilot were used to produce the final version of the assessment, which was administered as a large-scale assessment.

#### **First Cognitive Interview: Empirical Evidence 1 (EE1)**

**Results from cognitive interview with an expert.** A cognitive interview was conducted with an expert to verify that the intended reasoning will actually be enacted by a student if (s)he is at a certain level of IRS. Seventeen out of 35 items in AIRS-1 were asked to examine the expert's enacted reasoning. These 17 items were: (a) the items revised from the preliminary version of AIRS based on the experts' reviews; and (b) the items requiring high cognitive skills. It appeared that for all 17 items, the experts' verbal

reasoning matched well enough with the reasoning statement (intended reasoning). Table 14 presents some examples of the interview excerpts. The first three columns present the item number with the problem context, intended reasoning and the enacted reasoning (verbal script of the expert). The last column of the table presents the author's argument for why the expert's enacted reasoning was considered to be aligned with the intended reasoning. The reasoning statement and the expert's enacted reasoning for all 17 items are presented in Appendix K.

**Results from the first cognitive interview for item revision.** Item revisions were conducted based on results from the first cognitive interview with three students. Item revisions were made mostly to change the CR items to MC items. The response choices were constructed based on variations of the students' reasoning. Some items were revised in wording, specifically for items that students asked for clarification. Students' responses were analyzed focusing on—how they interpreted a question and how they reached an answer.

Table 14

*Excerpts of Expert's In-depth Cognitive Interview: Selected Notes*

Items (AIRS-1)	Intended Reasoning	Interview Excerpts: Enacted Reasoning (expert's reasoning)	Argument of Alignment
Item 5 (Spinner problem set: Null model)	The null hypothesis is the one that will happen assuming the spinner is fair: each letter has an equal chance of a quarter if we repeat spinning this spinner.	Since we have 10 spins, and we want to have a probability model, and we want to count the number of B's, based on the set-up of the spinner, it looks like each letter has an equal probability of being chosen, and because it's fair. The probability model is gonna be based on the fair spinner. Each letter would have to have an equal probability. This is a fair spinner in the long run; the probability of each letter would come out to be about one quarter.	The expert recognizes that the null model is the probability model that represents the probability of each letter appearing in the long run. She also understands that the spinner has an equal probability of showing up if this spinner is fair.
Item 10 (A drug company problem set)	Invalid. We need to see in which group the <i>chunk</i> of people has less time to get relief. This statement focuses only on some of the data, not about the general tendency of the data. (Students are expected to see the data as aggregates, not as individual data).	This statement is not valid. Because it looks to me like...if you look at the overall shape of these data, the overall average of the old formula would be larger than the overall average of the new formula, which means that the new formula works better.	The expert understands data as aggregates, not focusing on some of the individual data. She also looks at the "overall shape" and the "overall mean" to compare the two different samples of data.
Item 12 (A drug company problem set)	Invalid. Although the sample sizes are different for two groups, we can make a conclusion because both sample sizes are fairly large.	That is not valid. Two groups were chosen randomly; the number of samples is fairly large, so I think we can make some conclusion on the comparison.	The expert's verbal reasoning is perfectly matched to the intended reasoning statement.

(cont.)

Items (AIRS-1)	Intended Reasoning	Interview Excerpts: Enacted Reasoning (expert's reasoning)	Argument of Alignment
<i>Table 14, cont.</i>			
Item 13 (Biology and Chemistry)	Since the sample size and a difference between two samples look the same, we need to look at the distribution of the two. Biology has a narrower distribution indicating that the difference between the two groups is more consistent (or reliable), so it has stronger evidence that there is a difference between the two groups.	In both of the boxplots, the boxes overlap quite significantly. And the tails also overlap. For the chemistry, there is same amount of variability between the two strategies. And for the biology, there are fewer variations than the chemistry for both strategies. So I would say the less variability means the scores are more consistent in Biology. Given that the difference between the two strategies is almost the same in the two groups (Biology and Chemistry), the less variability gives stronger evidence against the claim.	The expert recognizes that the smaller the variability, the more consistent the data are. In comparing the two samples, she further understands that the data with less variability have stronger evidence of difference between the two groups, given that the observed difference is similar.



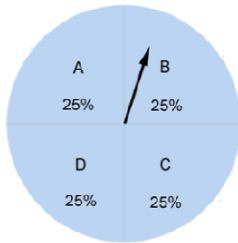
All six CR items were in the ISI part. For the first CR item in the Spinner problem set, “Which person do you think is correct and why?” the three students showed different reasoning. Student 1 answered, “I would say Person 2 is correct [5 Bs out of 10 spins is not unusual] because the sample size is not enough to say Person 1 is correct. We can’t say this is unusual.” The reasoning of student 2 was similar in that she also mentioned that the sample size was too small, but she chose C (Both are correct) because “there is no way to know which person is correct.” It is noted that students 1 and 2 chose different answers (B and C), but the reasoning behind their choices was the same. On the other hand, student 3 also chose answer C, but showed slightly different reasoning. She first considered the sampling distribution of statistics (the number of Bs in 10 spins) and then described where the observed sample statistic (5 Bs out of 10 spins) will be located in the distribution. She reasoned that each person is correct, offering a justification for each one. From the responses of student 2 and student 3, it is also noted that both chose answer C, but their justifications are different for why they thought both persons are correct.

It is debatable whether this item captures the original learning goal: being able to understand and articulate whether or not a particular sample of data is likely, given a particular expectation or claim. As seen above, the students’ reasoning did not match the intended reasoning behind the answer choice. More importantly, it appeared that each of the students showed reasonable justifications for their choices, indicating that all three response options are plausible. This indicates that this item is not properly assessing the learning goal, and that there are variations of correct reasoning that do not agree with the intended reasoning. Because of these issues, this item was removed. In terms of the

learning goal for this item, removing it did not affect the content coverage of the original test blueprint (items 3 and 5 assess similar learning goals).

For the other five CR items, alternatives were made from the students' responses. Question 5 shown below was originally an MC format item in the preliminary version, but it was changed to a CR format following the reviewers' comments, as described in Table 13 (not plausible alternatives). Students' answers about the null hypothesis were diverse, but all of the three students showed incomplete reasoning. Student 1 answered that the null hypothesis to test the fairness of the spinner, "5 or more B's out of 10 spins" and the alternative, "less than 5." A distractor was constructed from this *incorrect* reasoning: "The probability for letter B is  $1/2$  and the probabilities for the other letters sum to  $1/2$ ." Student 2 said, "The null would be that you would get 5 B's out of 10 spins, and the other letter would have the same spins," and another distractor was made from this reasoning: "The probability for letter B is  $1/2$  and the other three letters each have a probability of  $1/6$ ." Student 3 answered, "Five out of 10 could not happen just by chance," which was judged not to represent meaningful reasoning, and therefore, was not used to create a distractor for the MC format item.

**Questions 3 to 9 refer to the following:** Consider a spinner shown below that has the letters from *A* to *D*.



Let's say you used the spinner 10 times, and each time you wrote down the letter that the spinner lands on. Furthermore, let's say when you looked at the results, you saw that the letter *B* showed up 5 times out of the 10 spins.

Suppose a person is watching you play the game, and they say that it seems like you got too many *B*'s.

A second person says that 5 *B*'s would not be unusual for this spinner.

5. [Spinner problem set] A statistician wants to set up a probability model to examine how often the result of 5 *B*'s out of 10 spins could happen with the spinner just by chance alone. What would be the probability model the statistician can use to do a test? Please describe the null model.

A summary of student responses on each of the questions is presented in Table 15. Students' response choices are also shown. Incorporating the revisions made from the three think-aloud interviews resulted in the second version of the assessment (AIRS-2), which consisted of 34 MC items. Results from piloting AIRS-2 are discussed in the next section.

Table 15

*Excerpts of Students' 1st Cognitive Interview: Selected Notes*

Item	Student Reasoning in Think-alouds	Alternatives
<p>5. [Spinner problem set] A statistician wants to set up a probability model to examine how often the result of 5 B's out of 10 spins could happen with the spinner just by chance alone. What would be the probability model the statistician can use to do a test? Please describe the null model.</p>	<p>Student 1: "I am not exactly sure what the null model is. When it is the null hypothesis, <u>it will be 5 or more out of 10</u>; the alternative would be less than 5 out of 10."</p> <p>Student 2: "The null would be that you would get <u>5 B's out of 10 spins</u>, and the other letter would have the same spins. And the alternative [hypothesis] is that you would not get 5B's out of 10."</p> <p>Student 3: "A null model was the likelihood that something happens just by chance. The null hypothesis is kind of the opposite of the alternative hypothesis. The null hypothesis is that whatever you're suspecting is not true...I'm not being very clear. The null would be just the thing that did not happen. <u>The null hypothesis would be that five out of 10 could not happen just by chance.</u>"</p>	<p>a. The probability for each letter is the same—1/4 for each letter.</p> <p>b. The probability for letter B is 1/2 and the other three letters each have a probability of 1/6.</p> <p>c. The probability for letter B is 1/2 and the probabilities for the other letters sum to 1/2.</p>

(cont.)

Item	Student Reasoning in Think-alouds	Alternatives
<i>Table 15, cont.</i>		
6. [Spinner problem set] Are 5B's unusual or not unusual? Why?	<p>Student 1: "I do not think there is enough information because we do not have a small sample size. I guess 5 B's is unusual because it's supposed to be 25%."</p> <p>Student 2: "5B's are unusual. Because 5B's is in the tail; it didn't occur most often. A very low number happened."</p> <p>Student 3: "5 B's are unusual because it's well above the average number of (2 or 3) landing on B's."</p>	<p>a. 5 B's are not unusual because 5 or fewer B's happened in more than 90 samples out of 100.</p> <p>b. 5 B's are not unusual because 5 or more B's happened in four samples out of 100.</p> <p>c. 5 B's are unusual because 5B's happened in only three samples out of 100.</p> <p>d. 5 B's are unusual because 5 or more B's happened in only four samples out of 100.</p> <p>e. There is not enough information to decide if 5 B's are unusual or not.</p>
11. [Exam preparation problem set] ...Select either Biology or Chemistry and explain your choice.	<p>Student 1: "Chemistry. Because the boxplots are almost identical, and I see that the people in Biology, two groups (A and B strategies) look similar to each other. But in Chemistry, the range of strategy A is higher than B, so it does say that one strategy is better than the other." (faulty reasoning)</p> <p>Student 2: "First, I look at the ranges. The black lines are the medians, and it looks like both biology and chemistry are about the same. But biology has much narrower ranges. This means that the scores are closer together. So, I think biology."</p> <p>Student 3: "I think chemistry has the stronger evidence against the claim that neither strategy is better than the other. Because in Chemistry, somebody could argue that in chemistry somebody got almost 100 points for strategy A, but for strategy B, somebody only got 80 points. I guess for biology, you could do the same thing, but the range is bigger in Chemistry."</p>	<p>a. Biology, because scores from the Biology experiment are more consistent, which makes the difference between the strategies larger relative to the Chemistry experiment.</p> <p>b. Biology, because the outliers in the boxplot for strategy A from the Biology experiment indicate that there is more variability in scores for strategy A than for strategy B.</p> <p>c. Chemistry, because scores from the Chemistry experiment are more variable, indicating that there are more students who got scores above the mean in strategy B.</p> <p>d. Chemistry, because the difference between the maximum and the minimum scores is larger in the Chemistry experiment than in the Biology experiment.</p>

*(cont.)*

Item	Student Reasoning in Think-alouds	Alternatives
<i>Table 15, cont.</i>		
12. [Exam preparation problem set] ...Select either Psychology or Sociology and explain your choice.	<p>Student 1: "Sociology. Because it has a larger sample, but the other ones are the same; we could better believe that there is a difference."</p> <p>Student 2: "Psychology, because there is a lot variability in psychology. The smaller the sample size, the larger the variability."</p> <p>Student 3: "So it's the same type of question? So, sociology has a bigger sample size. Sociology has a smaller sample size, so it has more outliers. For sociology, it's clearer that every single line (outlier) in strategy B is higher than in strategy A. And that's also true for psychology, but the differences are less clear. This is also the same for Psychology, but in psychology, since it has a smaller sample size, we can't be so sure. Sociology has a larger sample, so it's more reliable."</p>	<p>a. Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment.</p> <p>b. Psychology, because there are more outliers in strategy B from the Psychology experiment, indicating that strategy B did not work well in that course.</p> <p>c. Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment.</p> <p>d. Sociology, because the sample size is larger in the Sociology experiment, which will produce a more accurate estimate of the difference between the two strategies.</p>

## Results from Pilot Testing: Empirical Evidence 2 (EE2)

**Analysis of pilot data.** The AIRS-2 was piloted to an introductory statistics course taught by a doctoral student in the summer of 2011. This assessment of 34 MC items was administered to 23 undergraduate students as a final exam. Students took the test online. The primary purpose of the pilot test was to identify potential deficiencies in the design, procedures, or specific items prior to a large-scale administration.

The mean for the total score was 23.26, with standard deviation of 4.93. A graphical representation of the distribution of the scores is presented in Figure 3. Item difficulties as a proportion correct are presented in Table 16.

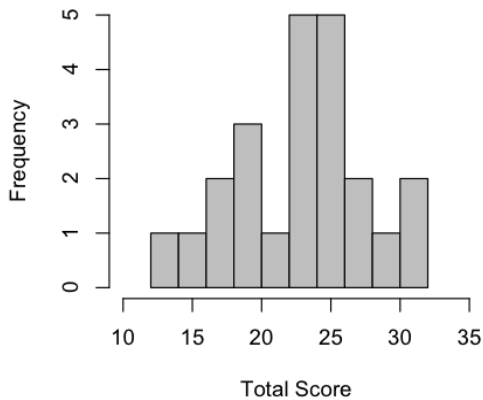


Figure 3. Distribution of total scores in pilot-test.

Table 16

*Item Difficulties (Proportion Correct) of AIRS Items*

Item	Proportion Correct	SD	Item	Proportion Correct	SD
1	0.43	0.51	18	0.78	0.42
2	0.87	0.34	19	0.7	0.47
3	1	0	20	0.65	0.49
4	0.96	0.21	21	0.96	0.21
5	0.61	0.5	22	0.87	0.34
6	0.22	0.6	23	0.57	0.51
7	0.65	0.49	24	0.91	0.29
8	0.87	0.34	25	0.22	0.42
9	1	0	26	0.57	0.51
10	0.87	0.34	27	0.52	0.51
11	0.74	0.45	28	0.39	0.5
12	0.48	0.51	29	0.87	0.34
13	0.87	0.34	30	0.78	0.42
14	0.35	0.49	31	0.91	0.29
15	0.57	0.51	32	0.65	0.49
16	0.48	0.51	33	0.65	0.49
17	0.87	0.34	34	0.43	0.51



Figure 4 displays the Q-Q plot to examine whether the distribution of the correct-total scores is normal. As seen in the plot, the distribution does not fundamentally depart from normality. The correct-total scores have a mean of 23.26 and a standard deviation of 4.93. Looking at the proportion correct (index of item easiness), it seems that item difficulties are distributed evenly across the 34 items. However, there are two items that all students answered correctly (item 3 and item 9), indicating these items may be easy and thus, may not perform well in discriminating students by ability.

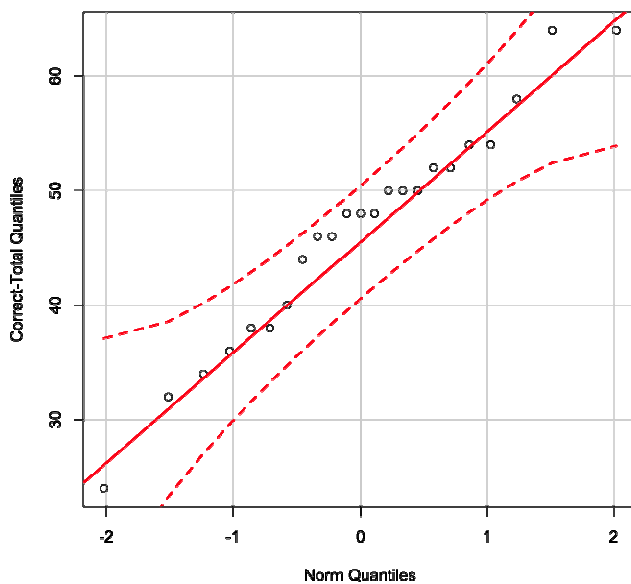


Figure 4. Q-Q plot of correct-total scores in pilot-test.

Both of these items are the first one in each of two scenarios, the Spinner scenario and the headache-medication scenario. Considering the learning goal for each item, as well as the logical sequence of the items within the set, both items were kept without any revision. However, the fact that these items are asked within a context gave rise to the issue of local dependency—each item in the same set does not provide unique

information regarding the students' level of IRS. If these two items are treated as one item in the testlet, the problem may be resolved since a testlet-score is produced by summing the scores for all items in a testlet.

The coefficient alpha for the pilot data was 0.84. As an indicator of strength of the relationship between the item score and total score, polyserial correlations based on tetrachoric correlations were obtained for each dichotomous item score (either 0 or 1). The correlations ranged from -.27 to 1. Results of a reliability coefficient analysis and polyserial correlations are shown in Appendix L.

There were three items with negative correlations between the item score and the correct-total score (item 4:  $r=-.27$ ; item 14;  $r=-.12$ ; item 29;  $r=-.14$ ). This indicated that these items do not function well in discriminating students who have high correct-total scores from those who have low correct-total scores. The author reviewed these items along with answer keys, item difficulties, and learning goals to investigate reasons for the negative item-total correlations. She decided to retain item 4 and item 29 without modifications, considering that the items (and alternatives) were carefully written to reflect students' reasoning during the cognitive interviews, and that these items are intended to measure important learning goals. Only item 14 was modified, which is shown in Table 17.

Table 17

*Changes Made in AIRS-3 from Pilot-testing*

Item in AIRS-2	Changes Made in AIRS-3 and Reason for the Change
<p>14. A random sample of 25 textbooks for different courses taught at a University is obtained, and the mean textbook price is computed for the sample. To determine the probability of finding another random sample of 25 textbooks with a mean more extreme than the one obtained from this random sample, you would need to refer to:</p> <p>a. the distribution of textbook prices for all courses at the University.</p> <p>b. the distribution of textbook prices for this sample of University textbooks.</p> <p>c. the distribution of mean textbook prices for all samples of size 25 from the University.</p>	<p>The sample size 25 was changed to 10.</p> <p>Option <i>a</i> is the distribution for the population of textbook prices. If we know this, it is reasonable to assume that we know the mean and SD for the population. Given that, we could approximate the distribution of sample means from random samples of size <math>n = 25</math> as <math>N(\mu, s/\sqrt{25})</math>. This is because with samples of size <math>n = 25</math> or larger, regardless of the shape of the population distribution, the distribution of sample means is approximately normal. In that sense, if we know <i>a</i>, we also know <i>c</i> (the distribution of mean textbook prices for all samples of size <math>n = 25</math>). If the sample size is small, there might not be a strong argument for <i>a</i>, and the best answer would be <i>c</i>.</p>

**Second Cognitive Interview: Empirical Evidence 3 (EE3)**

**Result of coding on think-aloud interviews.** This section presents the results of both the first and second cognitive interviews. There were three students in the first interview and six students in the second interview. A different item set was given to each student. Since there were six CR items (items 4, 5, 6, 7, 11, and 12) asked in the first interview, these items could not be coded into any of the four categories. Thus, these six items were not included in the coding process. Table 18 displays the coding results obtained from the first and second cognitive interviews. It includes counts of each code among four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP indicates that the interviewee selected a correct MC response option, and his (her) actual reasoning aligned with the intended reasoning. TN indicates

that the interviewee selected an incorrect MC response option, and his (her) actual reasoning was misaligned with the intended reasoning. FP indicates that that the interviewee selected a correct MC choice, but his (her) actual reasoning was incorrect. Finally, FN indicates that the interviewee selected an incorrect MC choice, but his (her) actual reasoning matched the intended reasoning.

The two categories, TP and TN, were considered to indicate “matched” in that these two codes indicate that a student’s response to an MC item matched the student’s actual reasoning. Similarly, FP and FN codes were considered to indicate “mismatched,” since a student’s MC response did not match the student’s actual reasoning. Table 18 presents the percentages of each category. Most of the items (30 out of 34) have a perfect match rate in terms of the relationship between the students’ actual reasoning and the MC response. These high rates provide evidence that a student’s score for each item represents the correctness of the student’s actual reasoning.

Table 18

*Coding Categories Made for Cognitive Interviews*

Item	# of Students Interviewed	Matched		Mismatched		Matched (%)	Mismatched (%)
		TP	TN	FP	FN		
1	6	2	4	0	0	100	0
2	6	5	1	0	0	100	0
3	7	7	0	0	0	100	0
4	5	3	2	0	0	100	0
5	5	4	1	0	0	100	0
6	5	2	2	1	0	80	20
7	3	2	1	0	0	100	0
8	7	5	2	0	0	100	0
9	2	2		0	0	100	0
10	4	4		0	0	100	0
11	2	1	1	0	0	100	0
12	4	2	2	0	0	100	0
13	7	6	1	0	0	100	0
14	4	4		0	0	100	0
15	2	2		0	0	100	0
16	2		2	0	0	100	0
17	4	3		0	0	100	0
18	5	2	3	0	0	100	0
19	6	3	3	0	0	100	0
20	7	3	2	2	0	71.4	28.6
21	5	5		0	0	100	0

*(cont.)*

Item	# of Students Interviewed	Matched		Mismatched		Matched (%)	Mismatched (%)
		TP	TN	FP	FN		
22	7	6	1	0	0	100	0
23	3	2	1	0	0	100	0
24	8	6	2	0	0	100	0
25	6	4	2	0	0	100	0
26	6	4		2	0	66.7	33.3
27	4	2	1	1	0	75	25
28	3	2	1	0	0	100	0
29	3	2	1	0	0	100	0
30	2	2		0	0	100	0
31	2	2		0	0	100	0
32	2	1	1	0	0	100	0
33	2	2		0	0	100	0
34	2		2	0	0	100	0

**Inter-rater reliability analysis.** Table 19 shows the results of coding for the interviews. The codes for 30 of the 34 items (88%) were aligned between the author and each rater. Cohen’s Kappa for the codes made on the two interview sets were 0.722 and 0.793, respectively. These values represent good inter-rater agreement, according to the cutoffs suggested by Landis & Koch (1977) and Altman (1991).

Table 19

*Results of Coding Cognitive Interviews*

	# of Item Total: 34	TP	TN	FP	FN
Alignment between the author and rater 1	Number of items agreed between the codes between author and rater 1	22	8	0	0
	Agreed total	30 (88%)		0 (0%)	
	Disagreed total	4 (12%); item 2; item 7; item 23; and item 26			
	Cohen's Kappa for 2 Raters (unweighted)	Kappa = 0.722 z = 4.45 p-value = 8.64e-06			
Alignment between the author and rater 2	Number of items agreed between the codes between author and rater 1	17	12	1	0
	Agreed total	30 (88%)			
	Disagreed total	4 (12%); item 12, item 16, item 18, item 20			
	Cohen's Kappa for 2 Raters (unweighted)	Kappa = 0.793 z = 5.93 p-value = 2.97e-09			

**Results from Field-testing: Empirical Evidence 4 (EE4)**

The analyses results for the responses obtained in a large-scale test administration are presented in the next three sub-sections: Local Item Dependence (LID), dimensionality, and IRT analysis. These analyses were based on the data collected from a large-scale administration with a representative sample of 1,978 students.

**Local Item Dependence (LID).** The presence of LID was investigated employing two methods: reliability analyses from the classical test theory (CTT) perspective and Chen and Thissen’s (1997)  $G^2$  local dependence (LD) indices from the item response theory (IRT) perspective.

**Reliability analyses.** Reliability of scores obtained from the 1,978 responses was evaluated using the CTT method. Two sets of reliability estimates are provided for each of the two forms of the response data: dichotomous data and testlet-based data. The Spearman-Brown formula was employed as a way to compare the reliability of the 34 discrete item responses and the reliability of testlet responses with respect to the effect of change of test length on the reliability (Sireci et al., 1991; Wainer, 1995). A summary of the coefficient-alpha and Spearman-Brown statistics is presented in Table 20.

Table 20

*Coefficient-alpha Reliabilities*

Test Format	Original Coefficient-alpha Reliability	Predicted Coefficient-alpha by Spearman-Brown Formula for 34 Items	Original Test Length
Dichotomous response data	0.805	0.805	34 items
Testlet-based data	0.771	0.857	19 items (9 testlet items and 10 discrete items)

The coefficient-alpha was lowered from .805 to .771, when the dichotomously scored data were aggregated into testlet-based data indicating the presence of LID.

However, as Sireci et al. (1991) pointed out, lower reliability in testlet-based data could



be due to the reduced test length. Therefore, the Spearman-Brown coefficient of reliability was employed to estimate the predicted reliability when the test length is increased by adding items with the same properties as those in the current test form. As shown in Table 20, the expected coefficient alpha is .857 when the number of items is increased to 34 from 19 in the original testlet test. The higher reliability for the testlet-based test confirms the overestimated size of the reliability in the 34-item test, and thus, the presence of LID. Moreover, this result suggests that use of testlet-based scoring provides more reliable and consistent score information.

*Local Dependence (LD) index.* For the 34 dichotomous items, Chen and Thissen's Local Dependence (LD) indices were examined. Fitting a unidimensional CFA model, an LD index matrix was obtained for each pair of 34 items (see Appendix M). Using this matrix, the mean of the absolute LD indices for each item with the other 33 items was computed (see Table 21). For the items given in a testlet, the LD index mean with the other items in the same testlet was computed. The LD index mean with the other items that are not in the testlet was also obtained for a comparison. As shown in Table 21, with respect to the testlet items, the mean LD indices for the testlet item pairs (row (b)) were quite large relative to the mean LD indices for all item pairs (row (a)), showing dependency of the items in the same context. This pattern becomes clearer when comparing the mean LD index between a testlet item and items in the same testlet (row (b)) to the mean LD index between the item and other items not in the same testlet (row (c)). Large differences in the magnitude of the mean indices between the two different circumstances (within the same passage or not) are evidence of the existence of LID. This

indicates that a pair or cluster of items in the same passage may need to be aggregated into a single unit-testlet.

Table 21

*Mean LD Indices of Each Item*

Item :Testlet												
Mean LD index	1	2	3: TL1	4: TL1	5: TL1	6: TL1	7: TL1	8: TL1	9: TL2	10: TL2	11: TL2	12: TL3
(a)With other 33 items	2.02	4.71	2.93	1.99	1.65	2.4	6.13	2.01	5.04	6.07	5.41	1.22
(b)With the items in the testlet			<b>4.75</b>	<b>4.25</b>	<b>2.6</b>	<b>2.92</b>	<b>5.24</b>	<b>4.82</b>	<b>15.15</b>	<b>52.55</b>	<b>43.83</b>	<b>6.23</b>
<sup>a</sup> (c)With the items not in the testlet			2.61	1.59	1.48	2.31	6.28	1.5	4.39	3.07	2.93	1.06

Item :Testlet												
Mean LD index	13: TL3	14	15: TL4	16: TL4	17	18	19: TL5	20: TL5	21: TL6	22: TL6	23	24: TL7
(a)With other 33 items	2.75	1.8	3.14	3.44	3.12	1.7	1.82	2.88	11.9	11.9	1.85	1.94
(b)With the items in the testlet	<b>6.23</b>		<b>56.1</b>	<b>56.1</b>			<b>4.94</b>	<b>4.94</b>	<b>311.9</b>	<b>311.9</b>		0.26
<sup>a</sup> (c)With the items not in the testlet	2.64		1.49	1.79			1.72	2.81	2.53	2.54		2.05

Item :Testlet												
Mean LD index	25: TL7	26: TL7	27: TL8	28: TL8	29: TL8	30: TL8	31	32	33	34		
(a)With other 33 items	3.97	2.7	5.31	5.23	1.76	2.88	3.3	2.71	1.99	3.3		
(b)With the items in the testlet	<b>19.3</b>	<b>19.3</b>	<b>33.4</b>	<b>5.23</b>	<b>2.73</b>	1.14						
<sup>a</sup> (c)With the items not in the testlet	2.98	1.63	2.51	2.55	1.67	2.6						

<sup>a</sup>This mean index was computed by averaging the LD indices between each item in a testlet and the items not in the testlet.

**Dimensionality.** To investigate the dimensionality of the item responses, Confirmatory Factor Analysis (CFA) was conducted using Mplus (Muthén & Muthén, 2010). Two factor models (a unidimensional model and a 2-factor model) were examined and compared. The factor structure of the 2-factor model was specified to reflect the hypothesized structure—one factor consisting of the items assessing ISI, and the other factor of the items measuring FSI.

In conducting CFA, weighted least squares with mean and variance adjustment (wlsmv) was used as an estimation method due to the fact that the responses are categorical (Muthén, DuToit, & Spisic, 1997). An assessment of model quality was based on the evaluation of parameter estimates (e.g., factor loadings, variances) and fit indices. In assessing the factor structure of a model, high and statistically significant factor loadings and a combination of fit indices were considered to comprehensively evaluate model fit and corroborate results (Hoyle, 1995; Thompson, 2004).

Table 22 displays the results of the unidimensional and two-factor model solutions of fitting two data sets (34 dichotomous item scores and 19 testlet-based scores per participant). For both models, all factor loadings were significant at  $\alpha = 0.05$ . Out of 34 items in the dichotomously scored data, 29 items had factor loadings above the specified .30 cutoff (McDonald, 1997). For the testlet-based data, 16 out of 19 testlets (including 10 discrete items) had factor loading greater than .30. This indicates that a high percentage of the variance in the responses for 29 items was explained by the model.

A summary of fit indices across the four factor models is shown in Table 23. The two-factor model was better fitted to data for both test formats. However, these measures are not much different than those for the unidimensional model. Moreover, the fit indices

indicated a moderate to good model fit for all four models, according to the cutoffs suggested by Hu and Bentler (1999; CFI and TLI more than .85) and Browne and Cudeck (1993; RMSEA index less than .05). Given that the presence of LID suggested using testlet-based scores, the results indicate that the item responses are unidimensional. Moreover, from the comparison of fit indices for the two models, it appeared that the change of the Chi-square from the unidimensional testlet model to the 2-factor model is not statistically significant ( $\Delta\chi^2(1) = 1.89, p = .169$ ), thus supporting the unidimensional model in terms of parsimony. Therefore, the assessment measures one unitary construct of IRS, as opposed to the hypothesized structure with two constructs, ISI and FSI.

Table 22

*Factor Loadings*

34 Dichotomous Item Response Data				19 Testlet-based Response Data			
Unidimensional Model		2-factor Model		Unidimensional Model		2-factor Model	
Item	Estimate (S.E.)	Factor	Estimate (S.E.)	Testlet	Estimate (S.E.)	Factor	Estimate (S.E.)
<b>1</b>	<b>0.253 (0.030)</b>	<b>ISI</b>	<b>0.263 (0.031)</b>	<b>1</b>	<b>0.258 (0.030)</b>	<b>ISI</b>	<b>0.260 (0.030)</b>
2	0.531 (0.025)		0.550 (0.026)	2	0.555 (0.025)		0.561 (0.025)
3	0.647 (0.029)		0.670 (0.030)	TL1-1	0.503 (0.021)		0.508 (0.021)
4	0.380 (0.031)		0.393 (0.032)				
5	0.339 (0.030)		0.351 (0.031)				
<b>6</b>	<b>0.236 (0.034)</b>		<b>0.242 (0.035)</b>				
7	0.436 (0.026)		0.453 (0.027)	TL1-2	0.581 (0.021)		0.587 (0.021)
8	0.599 (0.022)		0.622 (0.023)				
9	0.730 (0.021)		0.758 (0.021)	TL2	0.645 (0.019)		0.652 (0.019)
10	0.555 (0.027)		0.576 (0.028)				
11	0.608 (0.022)		0.630 (0.023)				

*(cont.)*

34 Dichotomous Item Response Data				19 Testlet-based Response Data			
Unidimensional Model		2-factor Model		Unidimensional Model		2-factor Model	
Item	Estimate (S.E.)	Factor	Estimate (S.E.)	Testlet	Estimate (S.E.)	Factor	Estimate (S.E.)
<i>Table 22, cont.</i>							
12	0.302 (0.029)		0.311 (0.030)	TL3	0.449 (0.023)		0.453 (0.023)
13	0.466 (0.025)		0.482 (0.026)				
14	0.454 (0.026)	FSI	0.462 (0.027)	14	0.465 (0.027)	FSI	0.467 (0.027)
<b>15</b>	<b>0.136 (0.031)</b>		<b>0.139 (0.031)</b>	TL4	<b>0.271 (0.026)</b>		<b>0.272 (0.026)</b>
16	0.332 (0.028)		0.338 (0.029)				
17	0.703 (0.021)		0.715 (0.021)	17	0.710 (0.022)		0.713 (0.022)
18	0.499 (0.026)		0.507 (0.026)	18	0.517 (0.026)		0.519 (0.026)
19	0.364 (0.028)		0.370 (0.029)	TL5	0.463 (0.023)		0.465 (0.023)
20	0.384 (0.027)		0.390 (0.027)				
21	0.688 (0.027)		0.703 (0.027)	TL6	0.321 (0.028)		0.322 (0.028)
22	0.336 (0.032)		0.344 (0.032)				
23	0.367 (0.027)		0.371 (0.028)	23	0.377 (0.027)		0.378 (0.028)

*(cont.)*

34 Dichotomous Item Response Data				19 Testlet-based Response Data			
Unidimensional Model		2-factor Model		Unidimensional Model		2-factor Model	
Item	Estimate (S.E.)	Factor	Estimate (S.E.)	Testlet	Estimate (S.E.)	Factor	Estimate (S.E.)
<i>Table 22, cont.</i>							
24	0.522 (0.024)		0.531 (0.024)	TL7	0.447 (0.022)		0.448 (0.022)
<b>25</b>	<b>0.080 (0.032)</b>		<b>0.081 (0.032)</b>				
26	0.391 (0.026)		0.398 (0.027)				
27	0.530 (0.024)		0.540 (0.024)	TL8	0.613 (0.018)		0.616 (0.018)
28	0.369 (0.027)		0.376 (0.027)				
29	0.510 (0.024)		0.519 (0.024)				
30	0.459 (0.025)		0.466 (0.026)				
31	0.726 (0.019)		0.741 (0.019)	31	0.735 (0.020)		0.738 (0.020)
32	0.401 (0.027)		0.409 (0.027)	32	0.414 (0.027)		0.415 (0.027)
33	0.433 (0.026)		0.439 (0.027)	33	0.447 (0.026)		0.448 (0.027)
<b>34</b>	<b>0.158 (0.040)</b>		<b>0.161 (0.041)</b>	<b>34</b>	<b>0.189 (0.041)</b>		<b>0.190 (0.041)</b>

*Note.* The bold fonts indicate items with factor loadings of less than 0.3.



Table 23

*Fit Indices for Factor Models*

Data Format	Model (N=1,978)	Fit Index				
		Chi-square (df; P-value)	TLI	CFI	RMSEA (90% CI)	WRMR
Dichotomous response data	Unidimensional model	2492.979 (527, <0.001)	0.837	0.847	0.043 (0.041, 0.045)	1.961
	2-factor model	2400.960 (526, <0.001)	0.841	0.851	0.042 (0.041, 0.044)	1.940
Testlet-based data	Unidimensional model	472.742 (135, <0.001)	0.953	0.958	0.033 (0.029, 0.036)	1.337
	2-factor model	470.883 (134, <0.001)	0.958	0.966	0.033 (0.028, 0.036)	1.334

As a result, the 19 testlet-based response data were used in the remaining analyses (examining item properties, item- and test-information functions using item response theory), which are described next.

**IRT model for polytomous data and assumptions.** This section presents the results of fitting response data to an IRT model to evaluate item properties, item- and test information. The assumptions for applying an IRT model are first examined. The results of fitting the graded response model (GRM, Samejima, 1969) are described next.

*The assumptions for IRT models.* The major two assumptions in applying an IRT model to response data are local independence and unidimensionality. Local independence in a test means that there is no relationship between examinee responses to different items after accounting for trait abilities measured by a test. IRT models are not robust to the violation of the local independence assumption. Since applying an IRT model to local dependence response data could cause serious problems (e.g., biased parameter estimates and overestimated test information (Yen, 1993)), it is important to check these assumptions before applying an IRT model.

Unidimensionality of a test indicates that a single latent trait is measured from the entire set of items. However, the latent traits measured in many performance assessments are very likely to be multidimensional, mainly due to various factors such as planned test construct structure, unintended nuisance or construct-irrelevant variances, and mixed item format. When unidimensional IRT models are employed to fit multidimensional data, several issues arise: biased IRT parameter estimates (de Ayala, 1994; 1995); threatening the validity of any inferences from the single ability estimate (Reckase,

1985); and biased results in the analysis of differential item functioning (DIF; Ackerman, 1992).

Given the evidence above indicating that the testlet-based response data are essentially unidimensional and that those data address the presence of LID in the discrete 34-item response data, the IRT assumptions were met.

*Item parameter estimates.* Table 24 shows the estimated item parameters and standard errors obtained by applying the GRM. In this model, discrimination parameters were allowed to be unconstrained for each item. The parameter estimates are under the usual IRT parameterization shown below:

$$\log\left(\frac{r_{ik}}{1 - r_{ik}}\right) = \beta_i (z - \beta_{ik}^*)$$

where  $\beta_{ik}^* = \frac{\beta_{ik}}{\beta_i}$  (Rizopoulos, 2012).

In the GRM model, score categories are separated by category boundaries: for cases where the testlet-based responses have five categories (resulting from combining 4 discrete items), the five score levels are separated by four category boundaries: the boundary between score level 1 and 2, 2 and 3, 3 and 4, 4 and 5, respectively. In the example of testlet 1.1 created from four items (item 3 to 6), each score level from 0 to 4 indicates the number of items correct, and category boundaries are used to determine the probability of passing the steps required to obtain a particular score level.

Table 24

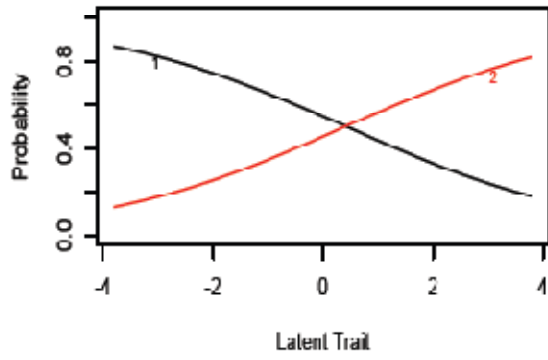
*Results of Fitting a GRM Model*

Item (or Testlet)	$\alpha$ (S.E.)	$\beta_1$ (S.E.)	$\beta_2$ (S.E.)	$\beta_3$ (S.E.)	$\beta_4$ (S.E.)
Q1	0.445 (0.055)	0.415 (0.116)			
Q2	1.142 (0.079)	0.250 (0.051)			
TL1.1	1.058 (0.059)	-3.505 (0.196)	-1.640 (0.155)	0.606 (0.110)	2.667 (0.655)
TL1.2	1.232 (0.071)	-1.053 (0.068)	0.557 (0.051)		
TL2	1.523 (0.084)	-2.236 (0.108)	-1.206 (0.108)	-0.202 (0.072)	
TL3	0.844 (0.058)	-0.875 (0.082)	1.754 (0.177)		
Q14	0.894 (0.069)	0.168 (0.060)			
TL4	0.475 (0.050)	-1.285 (0.166)	2.554 (0.321)		
Q17	1.923 (0.138)	-1.001 (0.055)			
Q18	1.045 (0.075)	0.434 (0.058)			
TL5	0.910 (0.061)	-2.202 (0.143)	0.525 (0.057)		
TL6	0.522 (0.062)	-4.252 (0.486)	-1.911 (0.269)		
Q23	0.691 (0.062)	-0.132 (0.073)			
TL7	0.889 (0.056)	-2.238 (0.142)	0.131 (0.057)	2.140 (0.296)	
TL8	1.387 (0.069)	-1.933 (0.092)	-0.678 (0.072)	0.394 (0.060)	1.519 (0.268)
Q31	2.060 (0.142)	-0.700 (0.044)			
Q32	0.757 (0.064)	0.383 (0.072)			
Q33	0.841 (0.069)	-0.689 (0.077)			
Q34	0.409 (0.073)	4.422 (0.760)			
Log-likelihood = -30946.84 AIC = 61999.68 BIC = 62295.94					

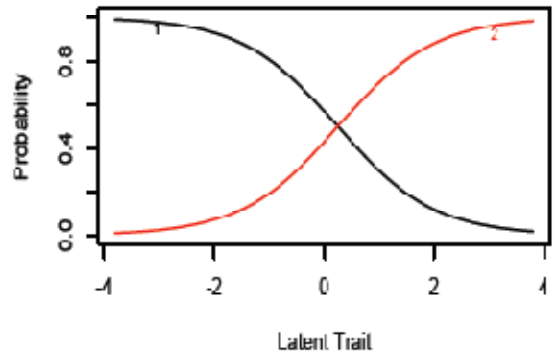
Table 24 also shows the estimates of item properties (item discrimination, thresholds between category boundaries) for 34 items. The items show acceptable discrimination capacity, and it appears that the instrument should perform well in estimating individuals in the approximate range of -2.5 to 2.5. The items (or testlets) have moderate to high discrimination estimates, ranging from 0.409 to 2.06, according to the qualitative classification proposed by Baker (1985; very low < 0.20, low = 0.21-0.40, moderate = 0.41-0.80, high > 0.80).

The location (difficulty) parameter  $b_i$  for each of the  $k$  category boundaries shows that the difficulty estimates are distributed evenly—from low to high. The patterns of a- and b-parameters are also represented in the Item Characteristic Curve (ICC) or Item Category Characteristic Curve for each testlet (see Figure 5). The ICC of each item is the plot of the probability as a function of theta for each category option.

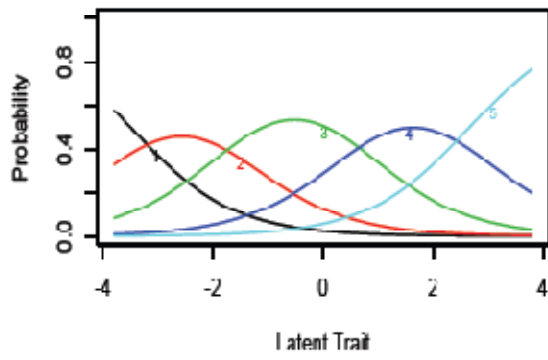
**Item 1**



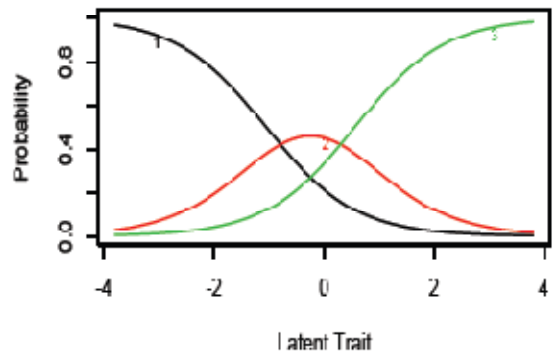
**Item 2**



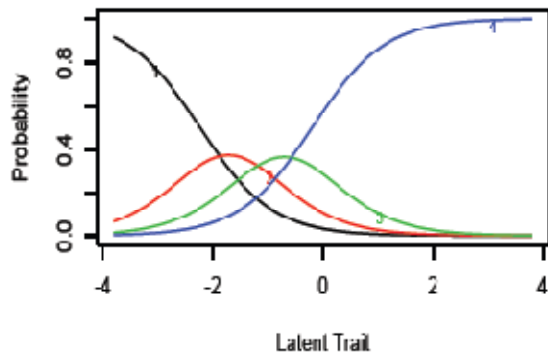
**Testlet 1.1**



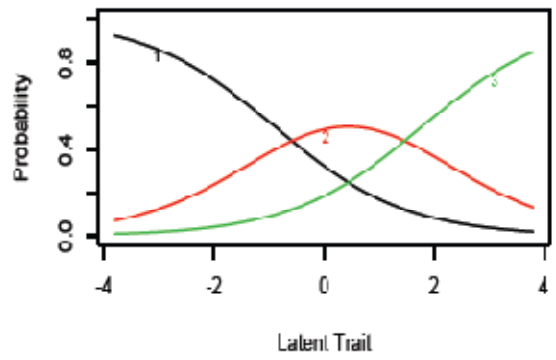
**Testlet 1.2**



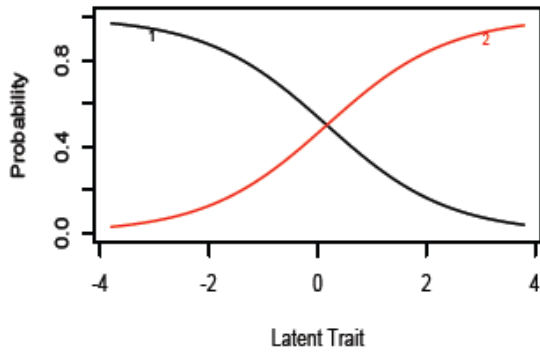
**Testlet 2**



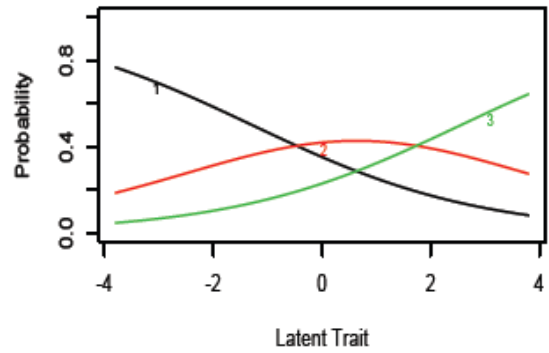
**Testlet 3**



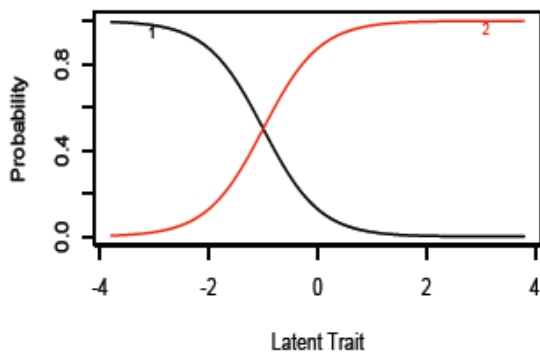
**Item 14**



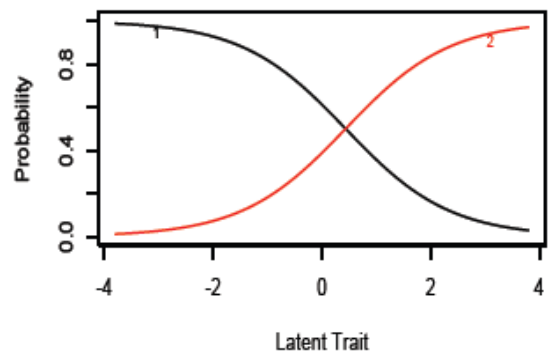
**Testlet 4**



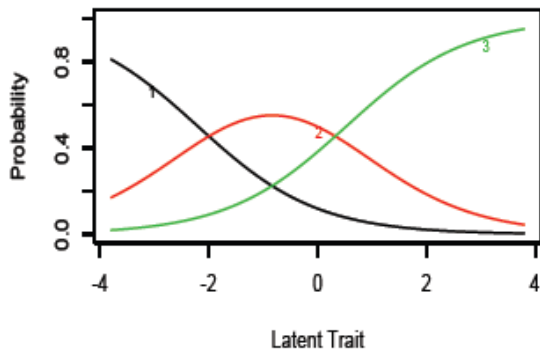
**Item 17**



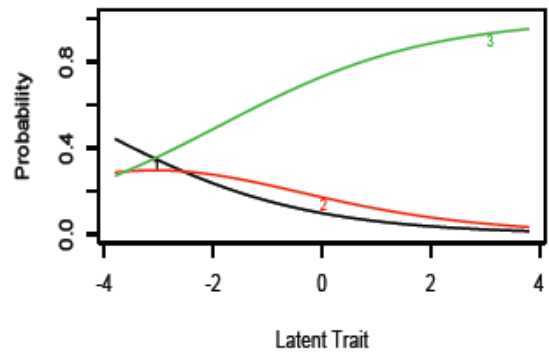
**Item 18**



**Testlet 5**



**Testlet 6**



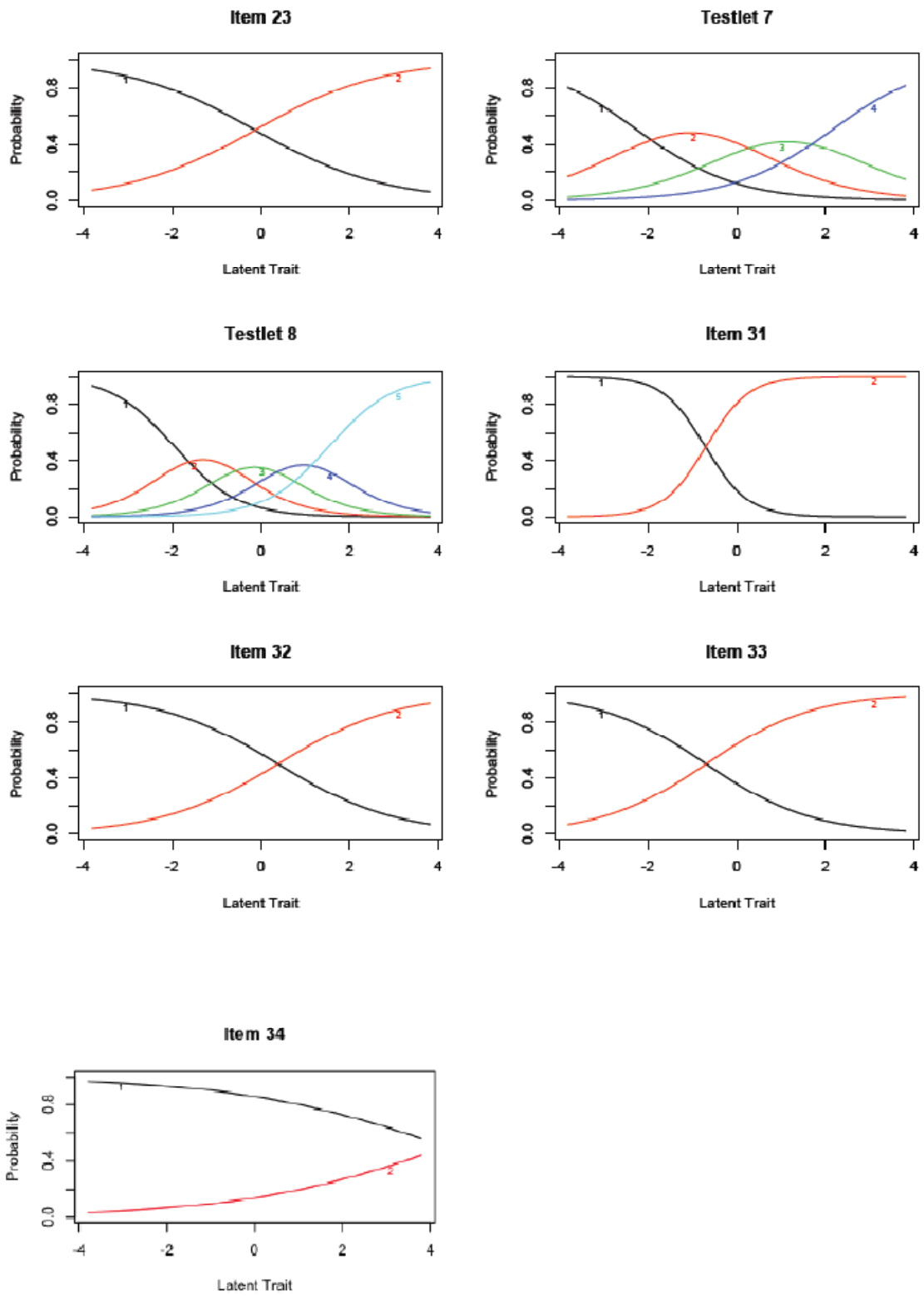


Figure 5. Item characteristic curves of 19 testlet-based items.



***Precision: Item information, Test information and Standard Error of***

***Measurement (SEM)***. Figure 6 displays the item information curves of the 19 testlet items. An item information curve is an index indicating the latent trait levels of IRS over which the item is most useful for distinguishing among individuals. Information curves with high peaks denote items with high discrimination, thus providing more information over the trait levels around the item's estimated thresholds. The information curves of item 1, testlet 4, testlet 6, and item 34 marked by dashed lines in Figure 6 show that these items have little precision in estimating trait levels.

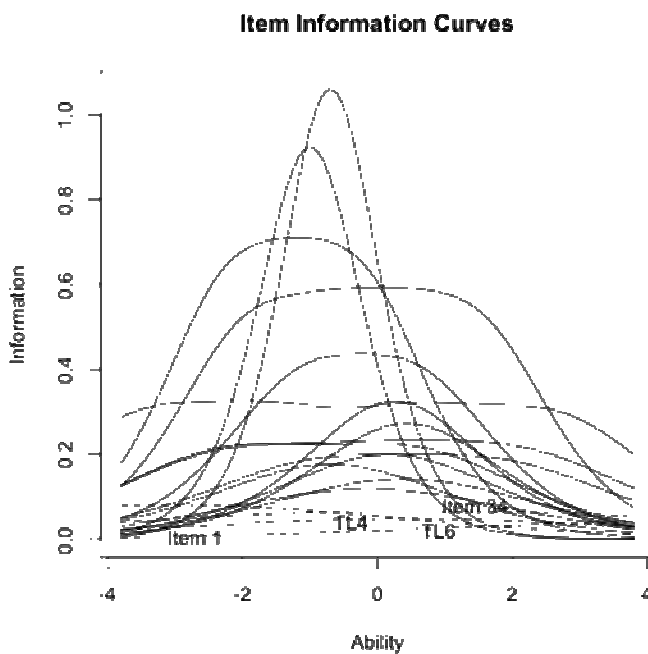


Figure 6. Item information curves of 19 testlet-based items.

In IRT, uncertainty about a person's location is quantified through the estimate's standard error of measurement (SEM). The SEM specifies the precision with respect to the person location parameter,  $\theta$ . From another perspective, test information is the

amount of information we have for estimating a person's location with an instrument, and it predicts the accuracy to which we can measure any value of the latent ability.

Therefore, there is a reciprocal relationship between SEM and test information, as represented below:

$$\text{Var}(\hat{\theta}) = \frac{1}{I(\hat{\theta})}, \text{ and thus,}$$

$$\text{SEM}(\theta) = \sqrt{\frac{1}{I(\theta)}}.$$

Figure 7 presents the information function of the test (based on the 19 testlet responses) and the SEM. It appears that the best precision for this test is for people with latent trait levels around zero. The standard error increases as the latent trait level gets higher (or lower), indicating that the items do not measure students who are above or below average very accurately.

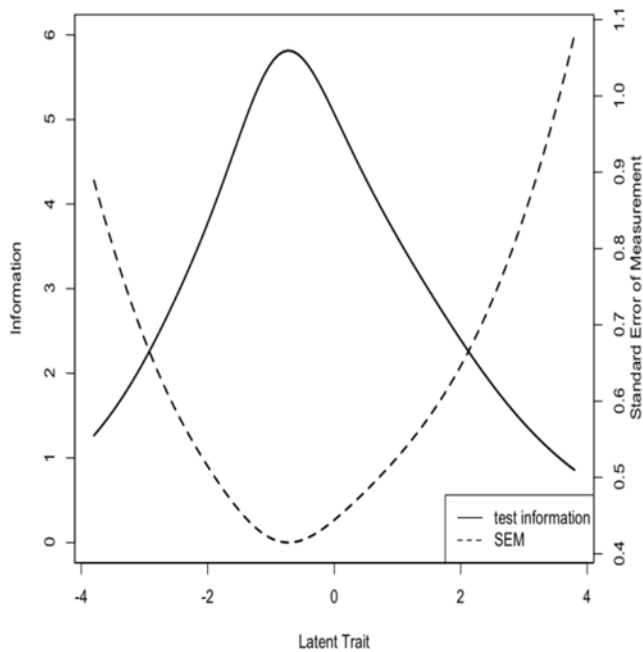


Figure 7. Test information function and standard error of measurement.

## **Synthesis of the Results**

This study sought to make multiple validity inferences to argue that scores derived from the AIRS test can be used to assess students' standing on the latent trait IRS in two content areas, ISI and FSI, and to provide information for a formative assessment in introductory statistics courses. Each inference in the interpretive argument prompted a particular investigation of the test development and evaluation procedures. Underlying inferences were evaluated by judging the claims laid out in the formative stage. Evidence sources collected in two stages were investigated to address the claims.

This section synthesizes the inferences to develop a validity argument narrative that captures the evolving evaluations of the test score interpretations and uses. The four inferences are revisited and critically examined. The theoretical evidence (TE1 to TE5) and empirical evidence (EE1 to EE4) served as resources to evaluate the plausibility of the claims.

### **Evaluation of Scoring Inference**

This inference is verified if Claim 3 (obtaining scores that are sufficiently precise) is supported. The following evidence resources were investigated to examine the plausibility of this claim: experts' judgments of the appropriateness of the answer key for each item, testing conditions, and scoring methods. Scores on the test obtained from CTT and IRT were examined and compared in terms of score precision. Item consistency (reliability) from a CTT perspective and item discrimination from an IRT perspective were examined.

During the experts' review of the preliminary assessment, an answer key was provided for each item. All three experts agreed to the answer key for each item. Since

the assessment items are all multiple-choice format, there is high confidence in the accuracy of the scoring, given that the items have only one best answer and that the scoring key is correct (Kane, 2004). However, there might be circumstances that can alter the interpretation of the scores. In field-testing, the testing conditions were different, depending on the institution and the instructor: there were some cases where the test was administered in a proctored environment by the instructor, and in other cases, students took the test in a convenient place (e.g., home or computer lab). There were also some variations in terms of use of the test scores; some instructors used the scores as part of their course grades, but others used the scores as extra credit. Different testing conditions might influence score accuracy; therefore, caution is needed in interpreting the test scores.

A distribution of the observed scores as number-correct is displayed in Figure 8. The mean of the testlet-based scores was 18.85 (N=1,978) with a standard deviation of 5.8. Figure 9 shows that the distribution of the observed scores as correct-total is approximately normal. The degree of precision for number-correct scores was based on reliability coefficients (coefficient-alpha) in CTT. In CTT, reliability coefficients (e.g., coefficient-alpha) are fixed for all scale scores (number-correct scores between 0 and 34), and in IRT, measures of score precision are estimated separately for each score level or response pattern, controlling for the characteristics (e.g., difficulty) of the items in the scale (Embretston & Reise, 2000). Test reliability has the advantages of being a very compact measure of precision. However, the most accurate estimates are those in which items are locally independent since item dependencies tend to inflate reliability estimation. When seemingly distinct items related to a context exhibit dependency,

grouping them together into a testlet more properly models the test structure (Sireci et al., 1991).

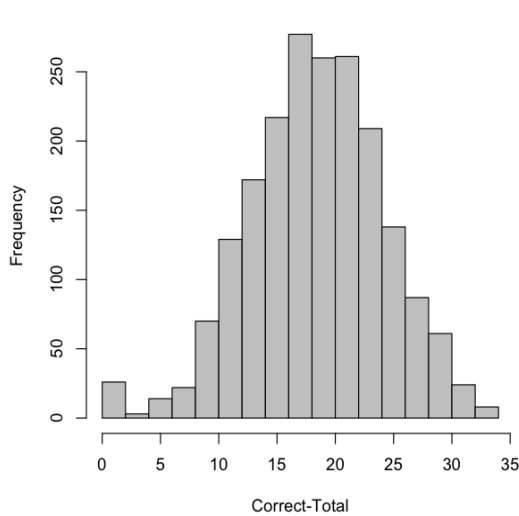


Figure 8. Distribution of correct-total scores (34-items total).

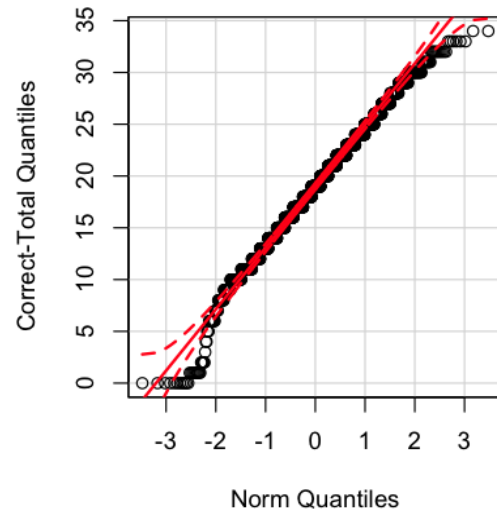


Figure 9. Q-Q plot of correct-total scores.

The reliability estimate obtained in EE4 was 0.81. This is above the recommended value of .70 suggested by Nunnally and Bernstein (1994). Since the coefficient alpha is a measure of internal consistency, calculated from the pairwise correlations between items, this level of reliability indicates that, on average, the items are measuring the construct of IRS consistently (precisely) at an acceptable level.

A distribution of the IRT-estimated scores on the latent trait is displayed in Figure 10. Figure 11 shows that the distribution of the ability levels is approximately normal. The mean of the estimates was -0.01 (N=1,978) with a standard deviation of 0.89. Item discrimination coefficients were examined to evaluate the scoring inference. The item discriminations shown in Table 24 in section 4.2.4 indicate that most of the items (or

testlets) have an appropriate level of discrimination (slopes in item characteristics curves) with moderate to high numerical values.

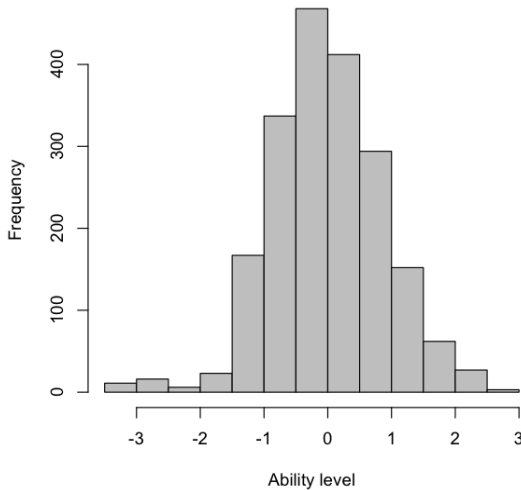


Figure 10. Distribution of IRT scores.

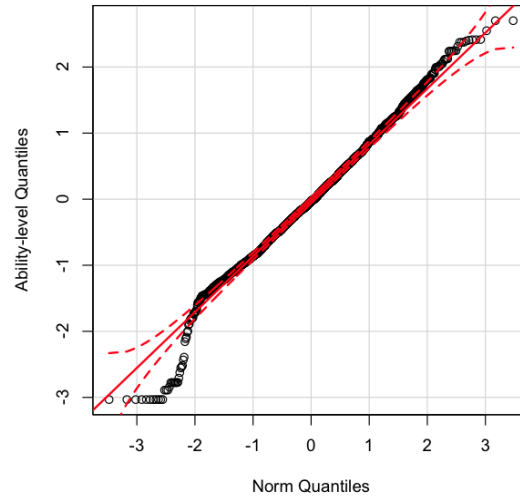


Figure 11. Q-Q plot of IRT scores.

However, an examination of item information curves suggests that item 1, testlet 4, testlet 6, and item 34 provide lower information relative to other items, indicating that they do not contribute much information in measuring the underlying trait. In other words, these items or testlets diminish the degree of score precision in measuring IRS.

Figure 12 shows a scatter plot of the scale scores graded by the GRM, plotted against the correct-total score (number-correct). This plot illustrates how advantageous IRT scoring methods are in dealing with scoring issues that may arise regarding score precision. One issue that may be questioned in the correct-total scores involves the use of summed “points” to score a test: why the rated “points” for the more discriminating items should be equal to the “points” for the less discriminating items. The IRT scale scoring process finesses this issue: all of the item responses are implicitly weighted; indeed, the

effect of each item response on the examinee’s score depends on the other item responses. Each response pattern is scored in a way that best uses the information about proficiency that the entire response pattern provides, assuming that the model summarizes the data accurately (Thissen & Wainer, 2001).

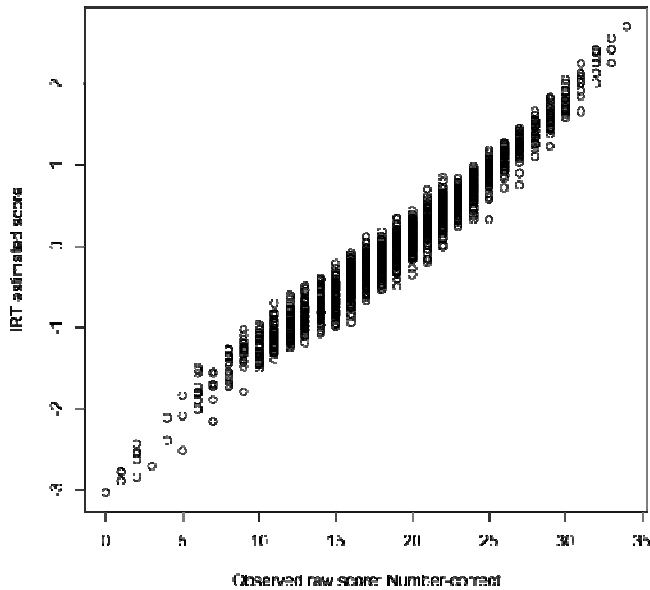


Figure 12. Scatter plot of correct-total scores (34 items) versus IRT scores.

As can be seen in Figure 12, the range of IRT scale scores is as much as a standard unit for some summed scores, although these scores are highly correlated ( $r=0.98$ ). For instance, the IRT scale score varied for examinees who obtained a summed score of 20 because some responded correctly to more of the highly discriminating items. Therefore, the IRT scale scores simultaneously provide more accurate estimates of each examinee’s proficiency and avoid any need for explicit consideration of the relative weights of the different kinds of “points.”

The evidence gathered throughout the assessment development procedure suggests that the AIRS test consistently measures the trait level of IRS *within* examinees, as shown in coefficient alphas and discrimination indices. When it comes to differences *between* examinees, however, a score is likely to be questioned in that the test administration conditions varied. Therefore, changing Claim 3 to reflect specific testing conditions (e.g., test proctoring, use of test scores) could better support the scoring inference in the validity argument.

**Evaluation of generalization inference (generalization from the score to the test domain).** Generalization inference concerns broadening the test score interpretation from an evaluation of a specific set of items to a claim about a student's expected score over the entire test domain (Kane, 2004). The plausibility of this inference was examined by asking the following question: To what extent do the test items and scoring represent the universe of generalization that is assessable from the target domain? This inference can be supported by evidence gathered for Claim 2, the test measures IRS in the representative test domains. In other words, evidence is needed to support the claim that tasks were sampled in a way to appropriately represent the range of tasks from the universe of generalization.

Four resources were used to explore the variance sources in generalizing from an observed score to a universe score: (a) construct representation documented in the test blueprint; (b) expert review of the test blueprint and the items; (c) cognitive interviews; and (d) standard error of measurement from item- and test-information.

The test blueprint documented the relevance of the test items to the learning goals by explicitly describing how each item is mapped to a specific learning goal that



represents the test domain (*Testing Standards*, 13.3). For example, in assessing the domain of “sampling variability,” item 2 measured the learning goal, “understanding the nature and behavior of sampling variability and taking into account sample size in association with sampling variability.” The degree of relevance between the test items and the learning goals documented in the test blueprint was evaluated by expert judgments.

In the expert review of the test items (TE5), three experts responded either “Strongly agree” or “disagree” to the evaluation question, “The items adequately assess the learning goals specified in each category.” One reviewer commented, “Knowing how difficult it is to write questions that assess statistical reasoning, I think that you have assembled some very good questions to assess your proposed learning goals. You have covered a wide range of situations using different types of data and methods (norm-based and randomization),” providing evidence of the congruency of the domain to measure and the test content. These results suggest that the test items properly cover the range of knowledge, concepts, and reasoning in the target domain of IRS.

Further, cognitive interviews using think-aloud provided evidence of how test scores represent their actual performance (reasoning) as indicators relevant to the broader domain (*Testing Standards*, 13.3). Matching two different measurement prompts, correct responses to MC items (1 or 0) and verbalizations of their reasoning, enabled evaluation of the extent to which generalization to the broader domain is supported. As shown in Table 18 in EE3 (Section 4.2.3), there were 30 items out of 34 that showed a 100% match between the correctness of MC choice (1 or 0) and alignment of student reasoning to the intended reasoning (aligned or misaligned), meaning that a student’s correct choice for an

MC item indicates the ability to make appropriate reasoning of the underlying content being assessed, and vice versa.

The inference from the observed score to the universe score was also explored using examinees' ability or trait parameters from the IRT analysis, although observed scores and trait parameters (universe scores) are stated in different units (AERA et al., 2002). An examination of the standard error of measurement played a major role in determining the precision of estimates of the expected score over the test domain; that is, the strength of the claim based on this estimate (Claim 2: To measure IRS in the appropriate domains; Brennan, 2001). The test information function summarized how well the test discriminates among individuals at various levels of the ability being assessed. The peak of the information curve of each item shown in Figure 5 (item information curves) indicated where on the theta continuum the test provides the greatest amount of precision, or information. As noticed, most of the items and testlets provided high information levels (i.e., less measurement error) somewhere around zero of the theta continuum and less information (i.e., high measurement error) as the theta goes to the extremes (-4 or +4). This pattern appears clearer in the test information function in Figure 6 showing that the SEM is higher as the theta level goes to either extreme.

Two potential sources of variability were identified as variability that prevents the generalizability inference. The first source of variability arises from an interaction between persons and items, coming from the educational and experiential histories that students bring to the performance, in this case, on the AIRS test (Shavelson & Webb, 1991). For example, the items asked in a Spinner context (items 3 to 8) would be easier for a student who has experienced a game using a spinner and who has thought about

probabilities in a fair spinner. The second source of variability comes from randomness, or other unidentified sources of variability (e.g., students took the test on different days, different testing conditions, etc.).

**Evaluation of extrapolation inference (extrapolating from the test domain to the IRS).** The tasks included in the AIRS test tend to be systematically different from the corresponding tasks in the domains of IRS (e.g., answering multiple-choice items about hypothesis testing is different from actual reasoning about hypothesis testing in a real context). The tasks in the test domain were de-contextualized versions of corresponding reasoning in the IRS domains. This inference regards extrapolation from performance on the test tasks to performance of the reasoning in the IRS domain (Kane, 2004). Three types of evidence were explored to verify this inference: expert review, think-aloud interviews, and dimensionality analysis.

The general evaluation form provided for the three experts included an evaluation question asking the extent to which the items measure students' IRS and not extraneous factors (e.g., test taking strategies or typical procedural knowledge). Two reviewers responded "agree" for this question, suggesting plausibility of Claim 1 (the test measures students' level of IRS) and Claim 5 (the test provides information about students' level of IRS).

The representativeness of the items in measuring IRS from reviewers' feedback was supported from cognitive interviews conducted with a graduate student and nine undergraduate students. Think-aloud data collected in one-on-one sessions where the candidates presented self-descriptions of how they approached each task provided a direct indication of how well a candidate's performance on each item of the test reflects

corresponding reasoning in IRS (Cronbach, 1971; Ohlsson, 1990). As revealed in the result of a think-aloud from a graduate student, the intended reasoning for all of the 34 items were actually elicited by the expert. This indicates that the expert's performance on the test reflected her reasoning on the corresponding items.

Another issue regarding the extrapolation inference is how the response data shows a structure of the test in terms of the hypothesized dimensionality (a single dimension of IRS or two dimensions represented by ISI and FSI). Given that the AIRS items were based on the test blueprint that reflects two content categories (ISI and FSI), separate scores from ISI and FSI domains could be obtained from the test if both theoretical, as well as empirical data, confidently support this structure. In an expert review of the test items, the review package included a form that asked about the extent to which the items distinguished between ISI and FSI. Two reviewers agreed that "the items reflect students' ISI or FSI" in general, and they also agreed that the items reflect the structure of ISI and FSI. However, an examination of dimensionality using confirmatory factor analysis revealed that the response data were closer to a unidimensional structure. This suggests that universe scores (IRT estimated scores) could provide inaccurate estimates if the scores were to be reported in two parts: one score for the ISI items and the other score for the FSI items. In other words, empirical evidence obtained from a large-scale administration shows that the students' estimated abilities represent (extrapolate) their level on one latent trait, IRS.

**Evaluation of explanation/implication inference.** Claims 4 and 5 concerned the extent to which AIRS test would help statistics instructors understand how students understand statistical inference, and give them useful information for a formative

assessment. To provide information for a formative assessment, it is necessary that the assessment covers multiple aspects of IRS (comprehensiveness of the test content) and that the test blueprint describing topics and learning goals helps instructors know what to look for when assessing IRS (a detailed and clear description of the blueprint).

Experts' positive evaluations provided during the blueprint and item review processes supported these arguments. The reviewers generally considered the blueprint as a good resource to be used as a framework in assessing statistical inference. As discussed in section 4.2, they acknowledged that the test blueprint covered multiple aspects of IRS. This was illustrated by reviewers' responses to the items: "The categories of the blueprint are well structured" (all rated "Agree") and that "the learning goals are clearly described" (one rated "Strongly agree" and two rated "Agree").

Given the agreement that the test can be functional to provide information in formative assessment measuring students' standing on IRS, the next question to be verified is how much information each item (as well as the test) provides in measuring IRS. Although the test provides a good amount of information across the latent trait levels, the standard errors of measurement (SEM) are high for students at low-ability and high-ability latent trait levels. This indicates that the test does not contribute as well to providing information for the students at these levels. It further suggests that a single observed score could provide an inaccurate estimate of a student's IRS proficiency in these ranges (high or low) of the latent trait.

## Chapter 5

### Summary and Discussion

This chapter summarizes the main research findings along with the discussion of the results and implications for teaching and for future research. Assumptions based on the validation results are discussed, as well as the extent to which the AIRS test scores provide useful and sufficient information for a formative assessment that measures inferential reasoning in statistics (IRS). Some of the claims are discussed focusing on discrepancies in results from theoretical evidence and empirical evidence.

#### **Summary of the Study**

This study developed and validated an assessment, the Assessment of Inferential Reasoning in Statistics (AIRS), designed to measure college students' inferential reasoning in statistics. The purpose of the assessment is to evaluate students' understanding of concepts of statistical inference in order to help statistics educators guide and monitor students' developing ideas of statistical inference.

Assessment development and validation were conducted by building and supporting arguments for the use of assessment in introductory statistics courses. In the two-phases of the research, the study first developed a test blueprint defining the target domains, and then developed the assessment from existing instruments and literature. Multiple sources of evidence were evaluated with regard to the plausibility of the inferences laid out from the test's claims.

In order for an observable attribute to be well defined, Kane (2006a) argues that the target domain must be clearly specified. The target domain in this study was defined in terms of the range of tasks (e.g., understanding sampling distributions, hypothesis

tests, evaluation of studies), test conditions (e.g., online test, 50- to 60-minute test), plausible contexts (e.g., classroom, home, or computer lab), and scoring rules (e.g., testlet-based scoring). Two content domains were specified from the literature—informal statistical inference (ISI) and formal statistical inference (FSI).

The *scoring* inference was supported through evidence regarding the appropriateness of scoring methods and precision of the scores. Use of a multiple-choice format provided high confidence in the accuracy of the scoring. During the expert review process, it was confirmed that all item answer keys were correct and that other responses were not debatable as alternative answers. Since the test responses showed the presence of local item dependence, testlet-based scoring was used.

During the item review process, the items were revised for clarification in wording, redundancy, and debatable issues. The observed scores showed an appropriate level of reliability in number-correct scores, but information provided from this score is limited in that there could be several students who have the same total-number-correct scores, but who would not be estimated to have the same latent trait level. The IRT estimated scores were used to address this issue since IRT considers the relative weights of the differential discrimination of each item. However, since testing conditions were different (e.g., taking the test at home, in a lab, or a classroom; different uses of the scores across courses), there should be some caution in interpreting the observed test scores, that is, in making an inference from an observed score to a universe score.

As Kane (2006a) argues, a *generalization* inference under the assumption of random sampling of tasks from the target domain is typically impossible to justify. Thus, it is more plausible to justify the claim that a set of tasks is representative of the universe

of generalization by evaluating if tasks were sampled in a way to appropriately represent the range of tasks from the universe of generalization. This was evaluated by examining that: (1) relevant topics and learning goals measured in each domain were included; and that (2) irrelevant tasks were absent from the test by confirming that no possible sources of bias were identified.

Expert reviews suggested that the items appropriately represent relevant topics and learning goals specified to measure the target domain of IRS. Results from student cognitive interviews confirmed that an observed score in the test represents a student's reasoning level on the latent trait. High correlation between observed scores (raw scores) and IRT estimated scores (universe scores) was another source of evidence supporting that an observed score in the test can be generalized to the score in the universe domain.

Students' estimated IRT scores represent their standing on the universe domain of IRS. It turned out that the IRT estimated scores were relatively precise and standard errors of measurement (SEM) were low in the range of -2 to 1 on the latent trait continuum. However, item information curves revealed that some items (items 1 and 34, and testlets 4 and 6) have low information functions (i.e., high SEM) suggesting the need for item revisions. Possible sources of variability, such as different testing conditions and students' familiarity with some items, could also reduce the magnitude of generalizability from an observed score to a universe score.

Evidence to support an *extrapolation* inference that a score in the universe domain can be extrapolated to the target domain was gathered by a think-aloud interview with an expert. The kinds of intended reasoning and skills required across the range of test tasks were elicited by the items, suggesting the skills being assessed in the tasks are



representative of those required to fully perform other tasks in the target domain. Results from a factor analysis suggested a unidimensional structure, providing evidence, to some extent, that the universe of generalization covers the target domain.

The inference regarding *implication/explanation* was examined using experts' qualitative reviews of the test blueprint and the test items. Positive evaluations about the comprehensiveness and clearness of the blueprint provided evidence that the test can be used to provide useful information for a formative assessment to understand student's current IRS. However, examination of item information functions revealed that there are some items that need to be improved in that those items contribute limited information in estimating student's current level of IRS.

### **Discussion of the Claims**

As reviewed in the literature, IRS has long been considered important, but difficult to develop (e.g., delMas et al., 1999a). In this regard, developing reasoning on ISI has been suggested as a "pathway" to help students learn and reason about formal concepts of statistical inference (e.g., Ben-Zvi, 2006; Makar & Rubin, 2009). If this conjecture that IRS involves two content domains, ISI and FSI, is empirically supported, this would provide educators and researchers with information to better develop students' current understanding of IRS.

In this study, there were claims made regarding the internal structure embedded in this test, and claims about test use and score interpretation drawn from the structure. Those claims are revisited below in terms of the plausibility based on theoretical evidence and empirical evidence.

## Is IRS Unidimensional or Multi-dimensional?

The following two claims were specified about the internal structure of the proposed test:

- Claim 1: The test measures students' level of IRS in two aspects—ISI and FSI.
- Claim 5: The test provides information about students' level of IRS in the aspects of ISI and FSI.

As it turned out, student's IRS as measured by this test did not support the hypothesized structure of two dimensions represented by ISI and FSI. There are a couple of plausible reasons for why the empirical data did not reflect a clear distinction between ISI and FSI. First of all, the two content domains of ISI and FSI are not clearly distinguished in the literature. Results from a factor analysis indicated that the response data were *essentially unidimensional* with a high correlation between the two domains.

Given that the items were designed as a two-dimensional structure and that the experts agreed that the items reflect this structure, the unidimensional result from response data suggest the following explanations of how students use ISI and FSI: A student who understands the ideas in FSI probably (1) uses FSI when it is required, (2) uses the ideas in FSI when only ISI is needed, or (3) uses both ideas in ISI and FSI when either are required. Considering that ISI is foundational to FSI, students with a good understanding of FSI might have a good understanding of ISI, and it may be that those who do not develop a good understanding of ISI have difficulty with developing FSI.

Pfannkuch's (2006b) perspective on statistical inference aligns to this result in that she views statistical inference as the ability to interconnect different ideas of

descriptive statistics as well as inferential statistics, within an empirical reasoning cycle. This implies that students might use both informal and formal methods of statistical inference even when they do not need to use formal statistical ideas. This further implies that students develop IRS as they interconnect different ideas and integrate them to generate appropriate reasoning processes. This aspect of IRS is also reflected in an argument suggested by Makar and Rubin (2009): inference is a multi-faceted construct.

### **How Useful is this Instrument?**

The following two claims are linked to the issue about uses of the proposed assessment.

- Claim 2: The test measures IRS in representative test domains.
- Claim 4: The test is functional for the purposes of formative assessment.

The test domains were specified based on a thorough literature review, and the test blueprint was developed laying out important topics and learning goals of each domain. Claim 2 was supported by experts' agreement that the topics and learning goals of the blueprint are comprehensive and the items well aligned to each item in the blueprint. This indicates that the AIRS can provide useful information for formative assessment (Claim 4).

In formative assessments, teachers evaluate student understanding of course materials to help them make better decisions in planning instruction. Teachers can then decide whether further review is required or if the students are ready for the introduction of new material (Thorndike, 2005). Given that Claim 2 was verified, teachers can refer to the test blueprint along with student response data on the AIRS test to identify content areas students find difficult to understand. In this way, teachers could use data from

student responses on this assessment for formative assessment and provide feedback to students to help them learn better.

### **Limitations**

While the results of this study supported the claims about the proposed test, there are some limitations that need to be considered. One of them concerns limited literature on the topic of inferential reasoning in statistics. Although inferential reasoning has been studied for decades, the study of statistical inference from teaching and learning perspectives is scarce. Due to the short history of statistics education as a discipline, there are no agreed upon definitions, content domains, and assessments to measure ISI and FSI as separate aspects. As seen in the blueprint- and assessment-review reports of the content experts, the reviewers had different opinions regarding the topics that need to be assessed. Although the author used the literature to decide which domains would be included, there are still arguable issues regarding what topics and learning goals are specifically about ISI and FSI.

Another limitation of the study is a lack of validity evidence based on relations to other variables (e.g., convergent and discriminant validity evidence). This study is missing this evidence source due to the nonexistence of a criterion measures to provide adequate comparisons. The generalization inference in the validity argument would be more strongly supported if there were evidence based on relationships with other variables as it addresses questions about the degree to which these relationships are consistent with the construct underlying the test interpretations (AERA et al., 2002).

Lastly, there are potential systematic sources of variability in test scores due to uncontrolled aspects of test administration. In the large-scale field-testing, instructors had

the flexibility to administer the online test depending on the course schedule, classroom environment, and student characteristics. This might result in lack of generalizability from the test score to the universe score.

### **Teaching Implications**

Although developing the concepts and ideas of IRS has been emphasized in teaching introductory statistics (ASA, 2005), many studies reported that students struggle with understanding formal concepts and procedures in inferential statistics (e.g., Haller and Krauss, 2002). Given that the students who participated in this large-scale assessment are representative of students enrolled in college-level introductory statistics courses, it would be worthwhile to look at the observed proportion-correct score (used as a measure of item difficulty) of each item or testlet to see in what areas college students show good understanding or difficulty. Here, the item difficulties were computed as a proportion-correct score from a CTT perspective instead of an IRT perspective since it is more straightforward in interpreting student's current level of understanding. Table 25 displays the item difficulties for each item or testlet.

Table 25

*Item Difficulties as Proportion-correct*

Items Asked Independently		Items Asked in Testlets			
Items	Item Difficulty	Items (Testlet)	Item Difficulty	Items	Item Difficulty
1	0.46	3 (TL1)	0.88 <sup>+</sup>	16 (TL4)	0.50
2	0.44	4 (TL1)	0.77 <sup>+</sup>	19 (TL5)	0.66
14	0.47	5(TL1)	0.37*	20 (TL5)	0.59
17	0.78 <sup>+</sup>	6 (TL1)	0.21*	21 (TL6)	0.87 <sup>+</sup>
18	0.41	7 (TL1)	0.50	22 (TL6)	0.75 <sup>+</sup>
23	0.52	8 (TL1)	0.61	24 (TL7)	0.64
31	0.71 <sup>+</sup>	9 (TL2)	0.82 <sup>+</sup>	25 (TL7)	0.35*
32	0.44	10 (TL2)	0.79 <sup>+</sup>	26 (TL7)	0.49
33	0.62	11 (TL2)	0.67	27 (TL8)	0.54
34	0.15*	12 (TL3)	0.34*	28 (TL8)	0.52
		13 (TL3)	0.53	29 (TL8)	0.54
		15 (TL4)	0.39*	30 (TL8)	0.53

\*: items with item difficulty less than 0.40

+: items with item difficulty greater than 0.70

Looking at the items with high proportion-correct, students seem to show good reasoning for items that asked either about a sample or a population separately. However, they tend to show incorrect reasoning if the items require them to connect reasoning about a given sample to a distribution of sample statistics and then to make a conclusion about a population.

For example, the two easiest items were items 3 and 4 shown in Appendix H.2. Items 3 and 4, which asked either for a particular sample or for a population as a separate question, had high proportion-correct scores. Even though they tend to show good understanding of how to set up a null model to examine whether a particular sample is unusual or not (item 4), many students didn't seem to understand what the null model represents in a distribution of sample statistics (item 5). They also showed lack of understanding of how to quantify unusualness and give a measure to argue that an observation is unusual (item 6).

The items with low proportion-correct (item 5 and 6) may indicate that students do not make a connection between an observed sample and the null model to make a conclusion about a population. To reason about this inference process correctly, students are expected to: (1) recognize what to support or reject (the null model), (2) find evidence from the observed results, (3) quantify the extent to which the evidence is unusual, and (4) make an argument for rejecting or not rejecting the null model based on the quantified measure of unusualness by going back to (1). This entire process was embedded in the set of items (question 3 to 8), and students were expected to use informal inferential reasoning to answer this set of questions.

Students' lack of ability to connect different ideas of IRS and unify them to make an appropriate conclusion is consistent with results from a study conducted by Makar and Rubin (2009). In characterizing students' informal statistical inference, these researchers found that students' initial attention to descriptive statistics (e.g., mean) for a sample never got back to the problem that would have allowed them to realize the potential of the data they collected as evidence for drawing inferences.

## **Implications for Future Research**

This assessment opened possibilities for future research about inferential reasoning in statistics. Further investigation is needed to use the AIRS from a longitudinal perspective in a classroom setting. The next step would be to observe students' assessment outcomes at different time points in a course, and to investigate how students' levels on IRS change over time as they learn formal inferential reasoning. This type of study could help track students' IRS from a developmental perspective so that students could be provided meaningful feedback.

There is also a need for more research studies to characterize the IIR associated with students' learning formal inference. It currently is not known how IIR is associated with IRS, how IIR affects IRS, and what instructional approaches are needed to develop IRS from IIR. There is a need for foundational studies about IIR to understand what kinds of informal ideas students have before they learn about formal concepts in statistics and how they use those ideas to learn about formal inferential ideas and techniques.

An improved assessment to measure students' IRS created in collaboration with statistics teachers and test developers would also be an interesting research area. The current practice of assessment design and development in introductory statistics courses is not well aligned with measurement or psychometric theories. Greater authenticity can result when test development is based on the joint consideration of content, item-quality and test-quality.

## **Conclusion**

Examination of multiple sources of evidence suggest: the newly created AIRS measures students' level of inferential reasoning in statistics (IRS) as a unidimensional



construct; the AIRS can provide useful information for formative assessment to understand students' current standing on IRS; and information obtained from the scores on this assessment is relatively precise and generalizable to a larger domain.

Incorporating these conclusions, it is suggested that this study contributes to the statistics education research in two ways: 1) This assessment will enable investigation of the impact of different approaches to teach the ideas of statistical inference using a reliable and valid measure; and 2) The AIRS provides a tool that can be used by instructors in statistics classrooms as well as by the statistics education research community. With the increasing attention being paid to effective way to teach statistical inference in introductory statistics courses these are two important contributions.

## References

- Aberson, C. L., Berger, D. E., Healy, M. R., Kyle, D. J., & Romero, V. L. (2000). Evaluation of an Interactive Tutorial for Teaching the Central Limit Theorem. *Teaching of Psychology, 27*, 289–291.
- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91.
- AERA, APA, NCME. (2002). *Standards for educational psychological testing*. Washington, DC: AERA.
- Altman, D. G. (1991). *Practical Statistics for Medical Research*. London, England: Chapman & Hall.
- American Statistical Association. (2005). *GAISE College Report*. Retrieved from ASA GAISE College Report Web site:  
<http://www.amstat.org/education/gaise/GAISECollege.htm>
- Aquilonius, B. C. (2005). *How do college students reason about hypothesis testing in introductory statistics courses*. Unpublished Ph.D. Thesis, University of California at Santa Barbara. Retrieved August 15, 2010, from  
<http://www.stat.auckland.ac.nz/~iase/publications/dissertations/05.Aquilonius.pdf>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.
- Baker, F. (1985). *The basic of item response theory*. Portsmouth, NH: Heinemann.
- Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy*,

- Reasoning, and Thinking* (pp. 147–168). Dordrecht, The Netherlands: Kluwer Academic.
- Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistics Education Research Journal*, 7(2), 130–145.
- Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, 2(1/2), 75–97.
- Batanero, C., Tauber, L. M., & Sanchez, V. (2004). Students' reasoning about the normal distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 257–276). Dordrecht, The Netherlands: Kluwer Academic.
- Beckman, M., Bjornsdottir, A., delMas, R., Everson, M., Garfield, J., Isaak, R., . . . & Zieffler, A. (2010). Evaluation report: Building a teaching and learning infrastructure. Evaluation conducted by the CATALST group at the University of Minnesota.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10, 389–396.
- Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*. [CD-ROM]. IASE, The Netherlands: International Statistical Institute. Retrieved August 17, 2010, from [http://www.auckland.ac.nz/~iase/publications/17/2D1\\_BENZ.pdf](http://www.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf)

- Ben-Zvi, D., & Garfield, J. (2011, July). New approaches to developing reasoning about samples and sampling in informal statistical inference. *The international collaboration of research on statistical reasoning, thinking and literacy (SRTL-7)*. Freudenthal Institute for Science and Mathematics Education, Utrecht University, Texel Island The Netherlands.
- Ben-Zvi, D., & Gil, E. (2010). The role of context in the development of students' informal inferential reasoning. *Proceedings of the Eighth International Conference on Teaching Statistics*. [CD-ROM]. IASE, Ljubljana, Slovenia, Invited Paper.
- Ben-Zvi, D., & Sharett-Amir, Y. (2005). How do primary school students begin to reason about distributions? In Reasoning about distributions: A collection of recent research studies. *Proceedings of the Fourth International Research Forum for Statistical Reasoning, Thinking, and Literacy (SRTL-4)*, Brisbane: University of Queensland, 2005. University of Auckland, New Zealand.
- Biehler, R. (2005). *Strengths and weaknesses in students' project work in exploratory data analysis*. Paper presented at the Fourth Congress of the European Society for Research in Mathematics Education, Sant Feliu de Guisols, Spain.
- Biggs, B., & Collis, F. (1982). Evaluating the quality of learning: The SOLO Taxonomy. New York, NY: Academic Press.
- Bond, T. G., & Fox, C. M. (2001). *Applying the rasch model (2nd)*. Mahwah, NJ: Lawrence Erlbaum.
- Brennan, R. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38, 295–317.

- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: SAGE.
- Carver, R. H. (2006, August). *Ambiguity intolerance: an impediment to inferential reasoning*. Paper presented at the Joint Statistics Meetings, Seattle, WA.
- Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48(3), 378–399.
- Chance, B. L., & Rossman, A. J. (2001). Sequencing topics in introductory statistics: A debate on what to teach when. *The American Statistician*, 55, 140–144.
- Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 1–25). New York, NY: Routledge.
- Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289.

- Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59(2).
- Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, 1(1).
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49(12), 997–1003.
- Collins, L., & Mittag, K. (2005). Effect of calculator technology on student achievement in introductory statistics. *Statistics Educational Research Journal*, 4(1), 7–15.
- Cox, D. R. (2005). Frequentist and Bayesian statistics: A critique (Keynote address). In *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, Oxford, England, 2005 (pp. 3–6). University of Oxford.
- Cronbach, L. J. (1971). Educational measurement. In R. L. Thorndike (Ed.), *Test validation* (pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1988). Five perspectives of the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–18). Hillsdale, NJ: Lawrence Erlbaum.
- Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools*, 5(2), 23–32.
- de Ayala, R. J. (1994). The influence of dimensionality on the graded response model. *Applied Psychological Measurement*, 18, 155–170.
- de Ayala, R. J. (1995). The influence of dimensionality on estimation in the partial credit model. *Educational and Psychological Measurement*, 55, 407–422.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

- delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, 4(1), 55–82.
- delMas, R. C., Garfield, J. B., & Chance, B. L. (1999). A model of classroom research in action: developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, 7(3).  
<http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm>
- delMas, R. C., Garfield, J.B., Ooms, A., & Chance, B. L. (2007). Assessing Students' Conceptual Understanding after a First Course in Statistics. *Statistics Educational Research Journal*, 6(2), 28–58. Retrieved September 10, 2010, from [http://www.stat.auckland.ac.nz/~iase/serj/SERJ6\(2\)\\_delMas.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf)
- delMas, R., Garfield, J., & Chance, B. (1999). Assessing the effects of a computer microworld on statistical reasoning. *Journal of Statistics Education*, 7(3).
- delMas, R., Garfield, J., & Chance, B. (1999). *Exploring the role of computer simulations in developing understanding of sampling distributions*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.
- Earley, M. (2001). *Improving statistics education through simulations – the case of the sampling distribution*. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, October 24–27.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (revised). Cambridge, MA: MIT Press.

- Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9*, 83–96.
- Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*, 75–98.
- Fennessy, L. M. (1995). *The impact of local dependencies on various IRT outcomes*. Ph.D. dissertation, University of Massachusetts at Amherst.
- Ferrara, S., Duncan, T., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., . . . & Westphalen, K. (2004). Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment. *Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.*
- Ferrara, S., Duncan, T., Perie, M., Freed, R., McGivern, J., & Chilukuri, R. (2003). Item construct validity: Early results from a study of the relationship between intended and actual cognitive demands in a middle school science assessment. Paper presented in S. Ferrara (Chair). *Cognitive and Other Influences on Responding to Science Test Items: What Is and What Can Be*. A symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*, 119–126.



- Fries, J., Bruce, B., & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experiment Rheumatology*, 23(5, Suppl. 39), S53–S57.
- Garfield, J. (1998). The Statistical Reasoning Assessment: Development and Validation of a Research Tool. In L. Pereira Mendoza (Ed.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 781–786). Voorburg, The Netherlands: International Statistical Institute.
- Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, 10(3).
- Garfield, J., & Ben-Zvi, D. (2008). *Developing Students Statistical Reasoning: Connecting Research and Teaching Practice*. Dordrecht, The Netherlands: Springer.
- Garfield, J., delMas, R., & Chance, B. (2002). ARTIST: Assessment Resource Tools for Improving Statistical Thinking, [Online]. [www.gen.umn.edu/artist/](http://www.gen.umn.edu/artist/)
- Garfield, J., delMas, R., & Zieffler, A. (in review). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM: The International Journal on Mathematics Education*.
- Garfield, J., delMas, R., & Zieffler, A. (2007). AIMS project. Retrieved from <http://www.tc.umn.edu/~aims/>
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

- Gigerenzer, G., Swijtink, A., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge, England: Cambridge University Press.
- Green, B. F., Bock, R. D., Humphreys, L. D., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347–360.
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1–20.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (pp. 65–110). Westport, CT: American Council on Education/Praeger.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd). Mahwah, NJ: Lawrence Erlbaum.
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, 7(1), 1–20.
- Halpin, P. F., & Stam, H. J. (2006). Inductive inference or inductive behavior: Fisher and Neyman-Pearson approaches to statistical testing in psychological research (1940–1960). *American Journal of Psychology*, 119(4), 625–653.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer/Nijhoff.
- Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12, 275–305.

- Hibbison, E. P. (1991). The ideal multiple choice question: A protocol analysis. *Forum for reading*, 22(2), 36–41.
- Hoekstra, R., Kiers, H., & Johnson, A. (2010). The influence of presentation on the interpretation of inferential results. *Proceedings of the Eighth International Conference on Teaching Statistics*. [CD-ROM]. IASE, Ljubljana, Slovenia.
- Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In *Data and context in statistics education: Towards an evidence-based society (ICOTS8)*, Voorburg, The Netherlands, 2010 (pp. CD-ROM). International Statistical Institute.
- Hong, E., & O'Neil, H. F., Jr. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology*, 84(2), 150–159.
- House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: SAGE.
- Hoyle, R. H. (1995). Structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 1–15). Thousand Oaks, CA: SAGE.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (a's) in classical statistical testing. *The American Statistician*, 57(3), 171–182.

- Innabi, H. (1999). Students' judgment of the validity of societal statistical generalization. In A. Roserson (Ed.), *Proceedings of the international conference on mathematics education into the 21<sup>st</sup> Century: Societal challenges, issues and approaches*.
- 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(4), 269–307.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kahneman, D., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*, New York, NY: Cambridge University Press.
- Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. *Proceedings of the Eighth International Conference on Teaching Statistics*. [CD-ROM]. IASE, Ljubljana, Slovenia, Refereed paper.
- Kane, M. T. (1992). An argument based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.
- Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21, 31–41.
- Kane, M. T. (2006a). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–154). Mahwah, NJ: Lawrence Erlbaum.

- Kane, M. T. (2006b). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). American Council on Education/Praeger.
- Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 19(2), 5–17.
- Kaplan, D. (2007). Computing and introductory statistics. *Technology Innovations in Teaching Statistics*, 1(1).
- Kaplan, J. J. (2009). Effect of belief bias on the development of undergraduate students' reasoning about inference. *Journal of Statistics Education*, 17(1). Retrieved from <http://www.amstat.org/publications/jse/v17n1/kaplan.html>
- Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61(2), 213–218.
- Konold, C. (1991). Understanding students' beliefs about probability. In E. V. Glaserfeld (Ed.), *Radical Constructivism in Mathematics Education* (pp. 139–156). Dordrecht: Kluwer Academic.
- Konold, C. (1994). Understanding probability and statistical inference through resampling. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings of the First Scientific Meeting (of the IASE)*; pp. 199–211). Perugia, Italy: Universita di Perugia.
- Konold, C. (2005). *Exploring data with TinkerPlots*. Emeryville, CA: Key Curriculum Press.
- Konold, C., Pollstek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education*, 24(5), 392–414.

- Krauss, S., & Wassner, C. (2002). How significance tests should be presented to avoid the typical misinterpretations. *Proceedings of the Sixth International Conference on Teaching Statistics*. [CD-ROM]. IASE, Cape Town, South Africa, Retrieved September 15, from <http://www.stats.org.uk/statistical-inference/KraussWassner2002.pdf>
- Kyburg, H. E. (1974). *The logical foundations of statistical inference*. Dordrecht, The Netherlands: Reidel.
- Landis, R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Lane-Getaz, S. (2007). *Development and validation of a research-based assessment: reasoning about P-values and statistical significance*. Ph.D. dissertation, University of Minnesota.
- Lane-Getaz, S. (2010). Linking the randomization test to reasoning about *P*-values and statistical significance. In Data and context in statistics education: Towards an evidence-based society, *Proceedings of the Eighth International Conference on Teaching Statistics*. [CD-ROM]. IASE, Ljubljana, Slovenia.
- Lane, D. M., & Tang, Z. (2000). Effectiveness of simulation training on transfer of statistical concepts. *Journal of Educational Computing Research*, 22(4), 383–396.
- Lavigne, N. C., Salkind, S. J., & Yan, J. (2008). Exploring college students' mental representations of inferential statistics. *Journal of Mathematical Behavior*, 27, 11–32.
- Lehman, E. L. (1991). *Testing statistical hypotheses*. New York, NY: Springer.

- Lipson, A. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. *Proceedings of the Sixth International Conference on Teaching Statistics*, Voorburg, The Netherlands, 2002. International Statistical Institute.
- Lipson, A. (2003). The role of the sampling distribution in understanding statistical inference. *Mathematical Educational Research Journal*, 15(3), 270–287.
- Lipson, K., Kokonis, S., & Francis, G. (2003). Investigation of students' experiences with a web-based computer simulation. *Proceedings of the 2003 IASE/ISI Satellite Conference on Statistics Education and the Internet*, Berlin [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute. Retrieved October 10, 2010, from <http://www.stat.auckland.ac.nz/~iase/publications/6/Lipson.pdf>
- Liu, Y. (2005). Teachers' understandings of probability and statistical inference and their implications for professional development. Unpublished Ph.D. Thesis. Retrieved from: <http://www.stat.auckland.ac.nz/~iase/publications/dissertations/05/liu.Dissertation.pdf>
- Liu, Y., & Thompson, P. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies*, 4(2), 129–138.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lunsford, M. L., Rowell, G. H., & Goodson-Espy, T. (2006). Classroom research: Assessment of student understanding of sampling distributions of means and the

- Central Limit Theorem in post-calculus probability and statistics classes. *Journal of Statistics Education* [On line], 14(3).
- Makar, K. (2009, July). The Role of Context and Evidence in Informal Inferential Reasoning. *The international collaboration of research on statistical reasoning, thinking and literacy (SRTL-6)*. University of Queensland, Brisbane, Australia.
- Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Educational Research Journal*, 8(1), 82–105. Retrieved from [http://www.stat.auckland.ac.nz/~iase/serj/SERJ8\(1\)\\_Makar\\_Rubin.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf)
- Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychology*, 34, 207–218.
- McDonald, R. (1997). Goodness of approximation in the linear model. In L.L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 199–219). Hillsdale, NJ: Erlbaum.
- Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction*, 14(2), 139–178.
- Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 391–423). Hillsdale, NJ: Erlbaum.
- Meletiou-Mavrotheris, M. (2004). Technological tools in the introductory statistics classroom: effects on student understanding of inferential statistics. *International*



- Journal of Computers for Mathematical Learning*, Dordrecht, 2004 (pp. 265–297). Kluwer Academic.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–104). New York, NY: American Council on Education/Praeger.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Metz, K. E. (1999). Why sampling works or why it can't: Ideas of young children engaged in research of their own design. In R. Hitt & M. Santos (Eds.), *Proceedings of the Twenty-First Annual Meeting of the North American Chapter of the International Group for the Psychology of Education* (pp. 492–498). Columbus, OH, 1999 ERIC Clearinghouse of Science, Mathematics, and Environmental Education.
- Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk*, *2*, 237–258.
- Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, *29*(4), 14–20.
- Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, *26*(1), 20–39.
- Moore, D. S. (2007). *The basic practice of statistics* (5<sup>th</sup> edition). New York, NY: W. H. Freeman.

- Moore, D., & McCabe, G. (2006). *Introduction to the practice of statistics* (4<sup>th</sup> ed.). New York, NY: Freeman.
- Moore, D., Notz, W., & Miller, J. (2008). Instructor's Manual and Test Bank with Solutions for *Statistics Concepts and Controversies* (7<sup>th</sup> ed.). New York, NY: Freeman.
- Muraki, E., & Lee, Y. (2001). *Detecting local item dependency in the TOEFL reading comprehension section: an application of the full-information item factor analysis*. Draft research report.
- Muthén, B. O., du Toit, 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; 6<sup>th</sup> ed.). Los Angeles, CA: Muthén and Muthén.
- National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics*. Reston, VA: NCTM.
- Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301.
- Noll, J. (2011). Graduate teaching assistants' statistical content knowledge of sampling. *Statistics Education Research Journal*, 10(2), 48–74.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

- Ohlsson, S. (1990). Trace analysis and spatial reasoning: An example of intensive cognitive diagnosis and its implications for testing. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Hillsdale, NJ: Erlbaum.
- Papariotodemou, E., & Meletious-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, 7(2), 83–106.
- Pfaff, T. J., & Weinberg, A. (2009). Do hands-on activities increase student understanding?: A case study. *Journal of Statistics Education*, 17(3). Retrieved from <http://www.amstat.org/publications/jse/v17n3/pfaff.html>.
- Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267–294). Dordrecht, The Netherlands: Kluwer Academic.
- Pfannkuch, M. (2006a). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, 5(2), 27–45.
- Pfannkuch, M. (2006b). Informal inferential reasoning. *Proceedings of the Seventh International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute. Retrieved from [http://www.stat.auckland.ac.nz/~iase/publications/17/6A2\\_PFAN.pdf](http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf)
- Pfannkuch, M. (2007, October). Year 11 students' informal inferential reasoning: A case study about the interpretation of box plots. *International Electronic Journal of Mathematical Education*. 2(3), 149–167.

- Phillips, L. D. (1973). *Bayesian statistics for social scientists*. London, England: Nelson.
- Pratt, D. (2007, August). Reasoning about Statistical Inference: Innovative Ways of Connecting Chance and Data. The international collaboration of research on statistical reasoning, thinking and literacy (SRTL-5). University of Warwick, England.
- Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, 7(2).
- Ramsey, F., & Schafer, D. (2002). *The statistical sleuth: A course in methods of data analysis* (2<sup>nd</sup> ed.). Belmont, CA: Duxbury Press.
- Reading, C. (2007, August). *Cognitive development of reasoning about inference*. Discussant reaction presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, England.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401–412.
- Reed-Rhods, T., Murphy, T. J., & Terry, R. (2006). The Statistics Concept Inventory: An Instrument for Assessing Student Understanding of Statistics Concepts", SIGMAA on Statistics Education session *First Steps for Implementing the Recommendations of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report*, Joint Mathematics Meetings, San Antonio, January 2006.
- Rizopoulos, D. (2012). Package 'ltm' (Version 09-7). Retrieved from <http://cran.r-project.org/web/packages/ltm/ltm.pdf>

- 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.
- Rosenthal, R. (1993). Cumulating evidence. In G. Keren (Ed.), *A handbook of data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276–1284.
- Rossmann, A. (2008). A statistician's view on the concept of inferential reasoning. *Statistics Education Research Journal, 7*(2), 5–19.
- Rossmann, A. (2008). Reasoning about informal statistical inference: one statistician's view. *Statistics Education Research Journal, 7*(2), 5–19. Retrieved from <http://www.stat.auckland.ac.nz/serj>
- Rossmann, A., Chance, B., Cobb, G., & Holcomb, J. (2008). CSI project. Retrieved from <http://statweb.calpoly.edu/csi/>
- Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (pp. 314–319). Dunedin, New Zealand: International Statistical Institute.
- Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. *Proceedings of the Seventh International Conference on Teaching Statistics*, Voorburg, The Netherlands: International Statistical Institute.

- Saldanha, L. (2004). "Is this sample unusual?": An investigation of students exploring connections between sampling distributions and statistical inference. Unpublished Ph.D. Thesis, Vanderbilt University.
- Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, 51, 257–270.
- Saldanha, L., & Thompson, P. (2006). Investigating statistical unusualness in the context of a resampling activity: students exploring connections between sampling distributions and statistical inference. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*. [CD-ROM]. IASE, The Netherlands: International Statistical Institute.
- Saldanha, L., & Thompson, P. (2007, October). Exploring connections between sampling distributions and statistical inference: An analysis of students' engagement and thinking in the context of instruction involving repeated sampling. *International Electronic Journal of Mathematics Education*, 2(3), 270–297.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* (17).
- Schafer, D. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education*, 61, 383–387.
- Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1(2), 115–129.

- Schmidt, F., & Hunter, J. (1997). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *Eight common but false objections to the discontinuation of significance testing in the analysis of research data*. Hillsdale, NJ: Erlbaum.
- Schneider, M., Huff, K., Egan, K., Tully, M., & Ferrara, S. (2010). Aligning achievement level descriptors to mapped item demands to enhance valid interpretations of scale scores and inform item development. In S. Ferrara & K. Huff (Chairs). *Cognition and Valid Inferences About student Achievement: Aligning Items with Cognitive and Proficiency Targets*. Cognition and Assessment SIG symposium conducted at the annual meeting of the American Educational Research Association, Denver, CO.
- Schwartz, D. L., Goldman, S. R., Vye, N. J., Barron, B. J., & the Cognition Technology Group at Vanderbilt (1998). In S. Lajoie (Ed.), *Aligning everyday and mathematical reasoning: The case of sampling assumptions* (pp. 233–273). Hillsdale, NJ: Erlbaum.
- Sedlemeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316.
- Sedlemeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavior Decision Making*, *10*, 33–51.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A Primer*. Newbury Park, CA: SAGE.
- Simon, J. L. (1976). Probability and statistics: Experimental results of a radically different teaching method. *The American Mathematical Monthly*, *83*(9), 733–739.

- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36, 477–481.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237–247.
- Smith, T. M. (2008). *An investigation into student understanding of statistical hypothesis testing*. Unpublished Ph.D. dissertation, University of Maryland.
- Sorto, M. A. (2006). Identifying content knowledge for teaching statistics. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*. [CD-ROM]. IASE, The Netherlands: International Statistical Institute.
- Sotos, A. E. C., Vanhoof, S., den Noortgate, W. V., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, 2, 98–113.
- Sowey, E. R. (2005). From a logical point of view: an illuminating perspective in teaching statistical inference. *International Journal of Mathematical Education in Science and Technology*, 36, 801–811.
- Stohl, H., & Tarr, J. E. (2002). Developing notions of inference using probability simulation tools. *Journal of Mathematical Behavior*, 21, 319–337.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality assessment. *Psychometrika*, 52(4), 589–617.
- Thissen, D., & Wainer, H. (2001). *Test Scoring* (1st). Mahwah, NJ: Routledge.



- Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, 26, 247–260.
- Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, 31, 113–123.
- Thompson, B. (1989). Asking “what if” questions about significance tests. *Measurement and Evaluation in Counseling and Development*, 61, 334–349.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25(2), 26–30.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- Thompson, P. (2004, April). *Why statistical inference is hard to understand*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Thompson, P., Liu, Y., & Saldanha, L. (2007). In M. Lovett & P. Shaw (Eds.), *Intricacies of statistical inference and teachers' understandings of them* (pp. 207–231). Mahwah, NJ: Erlbaum.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.

- Toulmin, S. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.
- Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*, 355–369.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105–110. (Reprinted in D. Kahneman, P. Slovic & A. Tversky [1982] *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.)
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 184*, 1124–1131.
- Upton, G., & Cook, I. (2008). Oxford Dictionary of Statistics (2nd). Retrieved October 12, 2010, from <http://www.oxfordreference.com.floyd.lib.umn.edu>.
- Vallecillos, A. (1999). Some empirical evidences on learning difficulties about testing hypotheses. In *Proceedings of the 52 session of the International Statistical Institute* (pp. 201–204). Helsinki: International Statistical Institute. Tome 58, Book 2.
- Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypotheses testing by university students. *Themes in Education, 3*(2), 183–198.
- Vallecillos, A., & Batanero, C. (1997). Conditional probability and the level of significance in tests of hypotheses. In *Proceedings of the 20th conference of the International Group for the Psychology of mathematics education*, Valencia, Spain, 1997 (pp. 271–278). University of Valencia.

- Vanhoof, S., Sotos, A., Onghena, P., & Verschaffel, L. (2007). Students' reasoning about sampling distribution before and after the sampling distribution activity. In *Proceedings of the 56 session of the International Statistical Institute*, Lisbon, Spain, International Statistical Institute.
- Wagner, D. A., & Gal, I. (1991). *Project STARC: Acquisition of statistical reasoning in children*. (Annual Report: Year 1, NSF Grant No. MDR90-50006). Philadelphia, PA: Literacy Research Center, University of Pennsylvania.
- Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Psychological Measurement*, 8(2), 157–187.
- Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, 32(7), 22–30.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15, 22–29.
- Watson, J., & Moritz, J. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education*, 31(1), 44–70.
- Watson, J. M. (2004). Chapter 12. Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 277–294). Dordrecht, The Netherlands: Kluwer Academic.
- Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior*, 19(1), 109–136.
- West, W. (2011). Textbooks 2.0. *Joint Statistical Meetings*, August 1, 2011.

- Well, A., Pollastek, A., & Boyce, S. (1990). Understanding of the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, 47, 289–312.
- Wild, C.K., Pfannkuch, M., Regan, M., & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference. *J. Royal Statistical Society A*. 174, Part 2, 1–23. Retrieved November 7, 2010, from [http://www.rss.org.uk/pdf/Wild\\_Oct.\\_2010.pdf](http://www.rss.org.uk/pdf/Wild_Oct._2010.pdf)
- Wilkerson, M., & Olson, J. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology*, 131(6), 627–631.
- Williams, A. M. (1999). Novice students' conceptual knowledge of statistical hypothesis testing. In J.M. Truran & K.M. Truran (Eds.), *Making the difference: Proceedings of the Twenty-second Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 554–560). Adelaide, South Australia: MERGA.
- Williams, A. M. (1999). Students' understanding of hypothesis testing: the case of the significance concept. In F. Biddulph & K. Carr (Eds.), *People in Mathematics Education: Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 585–591). Rotorua, New Zealand, MERGA.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58–79.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.

- Zenisky, A., Hambleton, R., & Sireci, S. (1999). *Effects of local item dependence on the validity of IRT item, test, and ability statistics*. Washington, DC: Association of American Medical Colleges.
- Zieffler, A., Garfield, J., delMas, R., Isaak, R., Ziegler, L., & Le, L. (2011). *How do tertiary students reason about samples and sampling in the context of a modeling and simulation approach to informal inference?* Paper presented at the Seventh International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-7). Texel Island, The Netherlands.
- Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Educational Research Journal*, 7(2), 40–58.
- 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 Minnesota.

## Appendix A

### Studies on Statistical Inference

Table A-1

*Studies on Foundations of Statistical Inference, Formal Statistical Inference, and Informal Statistical Inference*

Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
Aberson et al. (2000)	111	Undergraduates and graduates	Quasi-experimental study; Pre- and posttest	Quizzes and student ratings on their learning	Pre- and post-test: unclear Student rating: rating scale
Aquilonius (2005)	16	College students	1 group posttest	Classroom observation	
Bakker et al. (2008)	10	Employees	1 group posttest	Audio recordings, workplace artifacts, a questionnaire, and interview	
Batanero (2004)	117	Undergraduates	1 group; Pre- and posttest	Pretest: SRA (Konold & Garfield, 1993); posttest: questionnaire	Pretest: Multiple choice posttest: Open-ended questions

*(cont.)*

Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
<i>Table A-1, cont.</i>					
Belia et al. (2005)	473	Authors of journal articles	1 group one evaluation	Tasks presented in a website (Quantitative) Observations and interviews (Qualitative)	Open-ended tasks
Ben-Zvi (2006) [2]	75	Grade 5	1 group one evaluation	20 items from TIMSS	
Ben-Zvi & Gil (2010)	3	Grade 6	1 group one evaluation	Observation of students working on questionnaire (Qualitative)	Open-ended questions
Carver (2006)	48	College students	1 group; Pre- and posttest	Pre- and post-test (CAOS)	Multiple choice
Chance et al. (2004)	N=114 (pre- and post-test) N=37 (interview)	Undergraduates (pre- and post-test, interview) + graduates (interview)	1 group; pre- and posttest	Software (Sampling Distribution) Posttest and final exam	Multiple choice and interview
Collins & Mittag (2005)	22 versus 47	Undergraduates	Quasi-experimental: 2 groups	3 pretest scores; 1 inferential test scores; 1 final test score	Unclear
delMas & Garfield (1999)	49	Undergraduates	1 group posttest	1 posttest	Multiple-choice and true/false items
delMas et al. (1999) [2]	89 (initial activity); 141 (new activity)	Undergraduates	Quasi-experimental: 2 groups	Pre- and posttest	same items as delMas & Garfield, 1999

*(cont.)*

Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
<i>Table A-1, cont.</i>					
Earley (2001)	98	Undergraduates	1 group one evaluation	1 posttest	
Falk & Greenbaum (1995)	53	Undergraduates	1 group one evaluation	Questionnaire	One multiple-choice item
Grant & Nathan (2007)	3	Graduate students	1 group one evaluation	Interview	
Haller & Krauss (2002); Krauss & Wassner (2002)	44	Undergraduates	Quasi-experiment: 3 groups (instructors, scientists, and students)	Questionnaire	Six True/False questionnaires
Hertwig & Gigerenzer (1999)	18	Undergraduates	1 group one evaluation	Questions (interview)	Think aloud protocol
Hoekstra (2010)	71	Ph.D students	1 group one evaluation	Tasks on hypothesis testing and CIs	Open ended questions
Hong et al. (1992)	56	Graduate (N=27); Undergraduate (N=29)	Quasi-experiment: 4 experimental units	Pre-(10 items) and posttest (17 items)	A computer-assisted pretest; paper-and-pencil posttest
Kahneman & Tversky (1972)	95	Undergraduates	1 group one evaluation	Questionnaire	Open-ended
Kalinowski (2010)	94	Graduate students	1 group one evaluation	Survey	
Kaplan (2009)	10	Undergraduates	1 group one evaluation	Open-ended questions (interview)	

*(cont.)*



Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
<i>Table A-1, cont.</i>					
Konold et al. (1993)	88	16 high school students 25 undergraduates and 47 college students	1 group: pre and post-test	Open-ended questions	
Konold (1994)	199	High school students	1 group: pre and post-test	SRA (Konold & Garfield, 1993)	Multiple-choice
Lane-Getaz (2010)	105	College students in 3 introductory courses	Quasi-experiment: 3 groups: pre- and post-test	RPASS (Lane-Getaz, 2008)	34 Multiple-choice items
Lane & Tang (2000)	115	Undergraduates	Randomized control: Four treatments with two different conditions--factorial combination; And one control group	Pre- and post-test	12 open-ended questions
Lavigne et. al (2008)	3	Undergraduates	1 group evaluation (case study)	Word problem; Concept map; interview	Open-ended
Lipson (2003) [2]	23	Undergraduates	1 group one evaluation	Concept map	
Liu (2005); Liu & Thompson (2009)	8	High school mathematics teachers	1 group several evaluations (teaching experiment)	3 time interviews after each seminar	Video and interviews
Lunsford et al. (2006)	18 versus 7	Undergraduates	Quasi-experiment; Two groups	Pre- and post-test (27 items)	Items from delMas et al. (1999) (cont.)

Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
<i>Table A-1, cont.</i>					
Makar & Rubin (2007, 2009)	4	Primary school teachers (4) and their students (Grades unclear)	1 group one evaluation	Classroom observation and follow-up interview	
Means & Voss (1996)	60	Grades 5, 7, 9, and 11	1 group one evaluation	Interviews for Open-ended questions	
Meletiou-Mavrotheris (2004)	5	Undergraduates	1 group one evaluation	Experimental analysis (Videotape, classroom observation)	Transcript
Smith (2008)	104	Undergraduates	1 group one evaluation	Mixed methods: Assessment and follow-up interview (N=11)	14 multiple-choice item
Mittag & Thompson (2000)	225	AERA members (educational researchers)	Stratified random sample	Survey	
Paparistodemou & Meletious-Mavrotheris (2008)	22	Grade 3	1 group one evaluation (Case study)	Interview	
Pfaff & Weinberg (2009)	26	Undergraduates	1 group several evaluation	5 different assessments beginning/during/after instruction	Open-ended questions, items from delMas et al. (1999)
Pfannkuch (2005)	30	Grade 10	1 group one evaluation	Interview	

*(cont.)*

Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
<i>Table A-1, cont.</i>					
Pfannkuch (2006) [2]	1 teacher and 29 students	Grade 11	1 group one evaluation	Teacher communications and students communication	
Pratt (2008)	2	10- 11 years old	1 group one evaluation	Interview on students' working with activity	
Rubin et al. (1991)	12	Senior high school students	Observational	Interview of 6 open-ended questions	
Rubin et al. (2006)	9	Secondary Teachers (math/statistics)	1 group one evaluation	Responses to given tasks	
Saldanha (2004)	8	High school students	1 group one evaluation	Classroom observation; student written work	
Saldanha & Thompson (2003) [2]	27	High school students (11th and 12th grades)	1 group one evaluation	Classroom observation; student written work; post experiment interview	
Saldanha & Thompson (2006)	8	Grade 10 (N=1); Grade 11 (N=3); Grade 12 (N=4)	1 group one evaluation	Students discussion	

*(cont.)*

Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
<i>Table A-1, cont.</i>					
Sedlemeier (1998)	N=46 (Study 1) N=22+40 (Study 2) N=31 (Study 3)	Undergraduates	Study 1: random assignment of two conditions Study 2: extended interview from Study 1 Study 3: interview	Study 1: open-ended tasks in a PC Study 2: item with three tasks Study 3: interview	
Simon (1976)	25	Undergraduates	(Quasi-) Controlled experimental design	Pre- and posttest	Unclear
Sotos et al. (2009)	144	Undergraduates	1 group one evaluation	5 items from ARTIST project and confidence for the responses	Multiple choice (for assessment); 10-point Likert scale (confidence items)
Stohl & Tarr (2002)	2	Grade 6	1 group one evaluation (case study)	Analysis of students' work and conversation	
Thompson et al. (2007)	8	Teachers	1 group one evaluation	Seminar → Interview (Qualitative)	
Vallecillos (1995, 1996, 2000, 2002)	436	??	1 group one evaluation	Questionnaire and interview	20-item (true/false, multiple-choice, and open ended question)
Vallecillos and Batanero (1997)	7	University	1 group one evaluation	Questionnaire and interview	3 true/false items and two interview questions

*(cont.)*

Studies [number of studies]	Sample size	Students' Grade Level	Design	Instruments/Method of Data Collection	Response Methods
<i>Table A-1, cont.</i>					
Vanhoof et al. (2007)	221	Undergraduates	1 group pre- and post-test	Pre- and post-test	Multiple choice items during activity + 1 item from SRA
Watson (2004)	38	3 years after the previous study (Grades 6 to 13)	1 group repeated evaluations	Longitudinal interview with the same subjects	
Watson & Moritz (2000)	62	Grades 3, 6, and 9	1 group one evaluation	Interview and written works for open-ended questions	
Well et al. (1990)	1st study: N=114 2nd study a: N=151 2nd study b: N=138 3rd study: N=120	Undergraduates	1st study: 1 group 2nd study: 2 groups comparison 3rd study: groups comparison (controlled conditions)	1st study: questionnaires 2nd study: two versions of questionnaires for comparison 3rd study: problems for two groups, interview for 1 group	1st and 2nd study: two open ended questions; 3rd study: four open ended questions
Wilkerson & Olson (1997)	52	Graduates	1 group one evaluation	6 items	Type of items unclear
Williams (1999) [2]	18	Undergraduates	1 group one evaluation	Concept map and interview (pre- and post-interviews)	Talk aloud

Appendix B  
Preliminary Test Blueprint

Table B-1

*Test Blueprint to Assess Informal Statistical Inference*

Topic Category	Topics	Learning Goals	Literature	
207	Informal Inference (Inf-1)	The concept of uncertainty	Being able to express uncertainty in making inference using probabilistic (not deterministic) language	Makar and Rubin (2009), Zieffler et al. (2008)
	Inf-2	Properties of aggregates	Being able to able to reason about a collection of data from individual cases as an aggregate	Makar and Rubin (2009); Rubin, Hammerman, & Konold (2006); Pfannkuch (1999)
	Inf-3	Sampling variability	<ul style="list-style-type: none"> <li>- Understanding the nature and behavior of sampling variability</li> <li>- Understanding sample to sample variability</li> <li>- Taking into account sample size in association with sampling variability</li> </ul>	Rubin, Hammerman, & Konold (2006); Wild et al. (2011)
	Inf-4	The concept of unusualness	Being able to understand and articulate whether or not a particular sample of data is likely given a particular expectation or claim	Makar and Rubin (2009); Zieffler et al. (2008); Liu and Thompson (2009)
	Inf-5	Generalizing from a sample to a population	<ul style="list-style-type: none"> <li>- Being able to predict and reason about possible characteristics of a population based on a sample of data</li> <li>- Being able to draw a conclusion about population from sample(s) based on the prediction</li> </ul>	Zieffler et al. (2008)

*(cont.)*

Topic Category	Topics	Learning Goals	Literature
<i>Table B-1, cont.</i>			
Inf-6	Reasoning about comparison of two populations from two samples	<ul style="list-style-type: none"> <li>- Being able to predict and reason about possible differences between two populations based on observed differences between two samples of data</li> <li>- Being able to draw a conclusion about comparison of two populations from two samples based on the prediction</li> </ul>	Wild et al. (2011); Makar and Rubin, (2009); Zieffler et al. (2008); Pfannkuch, (2005)

Table B-2

*Test Blueprint to Assess Formal Statistical Inference*

Topic Category	Topics	Learning Goals	Misconceptions Found in Literature	Literature
Sampling distribution (SD-1) <sup>a</sup>	The concepts of samples and sampling	-Understanding the definition of sampling distribution -Understanding the role of sampling distribution	A tendency to predict sample outcomes based on causal analyses instead of statistical patterns in a collection of sample outcomes	Saldanha and Thompson (2002); Saldanha (2004); Rubin, Bruce, and Tenney (1991)
209 SD-2	Law of Large Numbers (Sample representativeness)	Understanding that the larger the sample, the closer the distribution of the sample is expected to be to the population distribution	A tendency to assume that a sample represents the population regardless of sample size ( <i>representativeness heuristic</i> )	Kahneman and Tversky; Rubin et al. (1991); Saldanha & Thompson (2002); Metz (1999); Watson & Moritz, (2000a, 2000b)
SD-3	Population distribution and frequency distributions	Understanding the relationship between frequency distribution and population distribution	Confusion between frequency distributions and sampling distributions	Sedlemeier (1997); Lipson, 2003; delMas et al. (1999)
SD-4	Population distribution and sampling distributions	Understanding the relationship between sampling distribution and population distribution	Confusion between population and sampling distributions	delMas et al. (1999)

*(cont.)*



Topic Category	Topics	Learning Goals	Misconceptions Found in Literature	Literature	
<i>Table B-2, cont.</i>					
SD-5	Central Limit Theorem	-Understanding the effect of sample size in sampling distributions -Understanding how sampling error is related to making an inference about a sample mean	Lack of taking into account sample size in association with distributions of samples	Mokros and Russell (1995); Sedlemeier & Gigerenzer (1997); Tversky & Kahneman, (1974); Vanhoof et al. (2007); Schwartz, Goldman, Vye, Barron, and The Cognition and Technology Group at Vanderbilt (1998); Wagner & Gal (1991); Well, Pollastek, and Boyce (1990)	
210	Hypothesis testing (HT-1) <sup>a</sup>	Definition, role, and logic of hypothesis testing	-Being able to describe the null hypothesis -Understanding the logic of a significance test	-Failing to reject the null is equivalent to demonstrating it to be true (Lack of understanding the conditional logic of significance tests) -Lack of understanding the role of hypothesis testing as a tool for making a decision	Batanero (2000); Nickerson (2000); Haller & Krauss (2002); Liu & Thompson (2009); Vallecillos (2002); Williams (1999); Mittag & Thompson, 2000
HT-2	Definitions of <i>P</i> -value and statistical significance	Being able to recognize a correct interpretation of a <i>P</i> -value	Misconception: <i>P</i> -value is the probability that the null hypothesis is true and that (1- <i>p</i> ) is the probability that the alternative hypothesis is true	Carver (1978); Falk & Greenbaum (1995); Nickerson (2000)	

(cont.)

Topic Category	Topics	Learning Goals	Misconceptions Found in Literature	Literature
<i>Table B-2, cont.</i>				
HT-3	<i>P</i> -value as a numerical probability	-Understanding the smaller the <i>P</i> -value, the stronger the evidence of a difference of effect -Understanding the relationship between <i>P</i> -value and standard error (Understanding that given the same mean difference, the smaller the variation in the sample statistic, the smaller the <i>P</i> -value, if all else remains the same)	Misconception: A small <i>P</i> -value means a treatment effect of large magnitude	Cohen (1994); Rosenthal (1993)
HT-4	Sample size and statistical significance in HT	-Understanding larger sample sizes yield smaller <i>P</i> -values, and more statistically significant observed results, if all else remains the same	Lack of understanding the relationship between sample size and statistical significance	Wilkerson and Olson (1997)
HT-5	Evaluation of HT	-Understanding that an experimental design with random assignment supports causal inference -Being able to make an appropriate conclusion from a hypothesis test	Lack of interpretation of result of hypothesis testing and statistical significance	Wilkerson & Olson (1997)
HT-6	Designing a statistical test for the comparison	-Being able to design a statistical test to compare two samples from a population -Being able to make a conclusion from a statistical test		

<sup>a</sup>SD and <sup>a</sup>HT: The SD was used to stand for the topic of *sampling distribution* and HT for the topic of *hypothesis tests*. However, in a later version of the blueprint, these acronyms were changed to *SampD* and *Stest* (See Appendix D), respectively. This is to avoid confusion that SD is used to represent *standard deviation* in statistics.

## Appendix C

### Expert Review Forms of Test Blueprint

#### **Consent Form: Expert Review**

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 validate the "*Assessment of Inferential Reasoning in Statistics (AIRS)*". You were selected as a possible participant because you have been contributing your expertise of college students' statistical reasoning and thinking on the research of the field of statistics education. 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: Jiyeon Park, Educational Psychology, EPSY 5261 instructor

#### **Background Information:**

The proposed study is to develop an instrument to assess two aspects of college students' statistical inferential reasoning—informal and formal statistical inference. The target population of the assessment is college students in the U.S. who are taking a non-calculus-based statistics course. The purposes of this assessment are: (1) to monitor students' longitudinal development of inferential reasoning as they learn statistics in an introductory course; and (2) to facilitate statistics education research on students' informal and formal statistical inference and the effect of instructional approaches on this topic.

#### **Procedures:**

If you agree to be in this study, we would ask you to take your time to review and evaluate the test blueprint and preliminary assessment on the evaluation form attached.

#### **Risks and Benefits of Being in the Study:**

There are no known risks to you as a participant.

The benefit to participation is the opportunity to contribute your expertise on the statistics education research.

#### **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. Research records will be kept in a locked file; 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 Jiyoan Park under the advisement of Professors Robert delMas, Ph.D. (Educational Psychology--Statistics Education) and Joan Garfield, Ph.D. (Educational Psychology—Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Jiyoan Park via my University of Minnesota, email: parkx666@umn.edu. You may also contact my advisor, Robert delMas, at delma001@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 can print 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.

You need to sign and return this consent form if you agree to let us use your responses in the research study described above.

I give permission for my responses to evaluation form to be included in any analyses, reports or research presentations made as part of this research project.

Your Name (Please PRINT):

\_\_\_\_\_

Signature \_\_\_\_\_

## The Invitation Letter and Test Blueprint Evaluation Form

Expert Invitation Letter March 22, 2011

Dear Professor XXX,

I am conducting my dissertation research on the development of an assessment to measure students' reasoning of statistical inference in two aspects—formal and informal inference. The purposes of the proposed assessment are: (1) to monitor college students' longitudinal development of inferential reasoning as they learn statistics in an introductory course, and (2) to facilitate statistics education research on students' informal and formal statistical inference and the effect of instructional approaches on this topic. With this letter I am formally soliciting your expert help in the development of my research instrument, which is now titled *Assessment of Inferential Reasoning in Statistics (AIRS)*.

As a sequential process of expert review in the development of the instrument, at the first stage, I am asking you to evaluate the test blueprint with respect to the validity of the topics and learning goals in the blueprint for developing an assessment to measure students' statistical inference. Please note that the learning goals that students have in reasoning about statistical inference, specifically in the two categories of informal and formal inference, were culled from research literature. As a statistics educator your expert opinion on how these items measure students' statistical inference is invaluable.

The assessment items will be developed from the test blueprint based on your feedback at the first stage. At the second stage, I will ask you to evaluate the assessment items that are developed from the test blueprint.

As an expert rater you are being asked to assess the validity of the blueprint and the assessment in relation to these specific learning objectives and misconceptions. If you are willing to participate in these two stages of expert review on the development of the instrument, please email me to confirm your interest at: [parkx666@umn.edu](mailto:parkx666@umn.edu).

I am attaching two documents to help you get a sense of the task I am asking you to perform: 1) the test blueprint, and 2) the evaluation form. The test blueprint is organized into two main sections, informal statistical inference and formal statistical inference. Formal statistical inference is categorized into two subtopics, sampling distributions and hypothesis testing. The evaluation form includes questions that ask about the validity of the content and the degree to which the test blueprint is relevant to the constructs, informal and formal inferential reasoning.

About 40 to 50 assessment items will be written based on the revised test blueprint. You will also be asked at a later time to rate each of the assessment items with respect to how well they measure the learning outcomes stated in the final test blueprint. You will be asked to suggest improvements for any items for which you “strongly disagree” or “disagree”. You will be asked to suggest concepts/topics that may be missing, items that can be removed/revised, and any other suggestions you may have to improve the assessment.

If you agree to participate as an expert reviewer, I will send you again a copy of the test blueprint for you to review. The turnaround for the evaluation form of the blueprint will be 2 weeks. Please feel free to ask me any questions that you have. I sincerely hope that you will be able to contribute to my research.

Thank you,

## Test Blueprint Evaluation Form

### Evaluation Form on the Test Blueprint

This is an evaluation form to get information of how valid the test blueprint is to develop an instrument to assess college students' informal and formal inference in statistics. Please read through the blueprint carefully before answering the items below.

Part 1. Please check the extent to which you agree or disagree with each of the following statements about the blueprint.

Item	Evaluation Questions	Ratings			
		Strongly agree ▼	Agree ▼	Disagree ▼	Strongly Disagree ▼
1	The topics of the blueprint represent the constructs of informal inference and formal inference in statistics.				
2	The learning goals of the blueprint are adequate for developing items to assess students' understanding of informal inference.				
3	The learning goals of the blueprint are adequate for developing items to assess students' understanding of formal inference.				
4	The set of learning goals is well supported by the literature.				
5	The learning goals are clearly described.				
6	The categories of the blueprint are well structured.				
7	The blueprint provides a framework for testing the constructs of informal and formal statistical inference.				

Part 2. For the following questions, please describe your opinions about the blueprint.

1. For each item to which you responded "Strongly disagree" or "Disagree", please explain why you disagree and suggest how the blueprint might be improved.
2. What do you think may be missing from the content of the blueprint related to the constructs of informal and formal statistical inference?
3. What parts of the blueprint may be extraneous or not as important for measuring the constructs of informal and formal statistical inference?
4. Do you have any other suggestions for improving the test blueprint? Please describe.

Thank you

## Appendix D

### Final Version Test Blueprint

Table D-1

*Test Blueprint to Assess Informal Inference*

Topic Category	Topics	Learning Goals	Items
Informal Inference (Inf-1)	The concept of uncertainty	Being able to reason about uncertainty in making inference using probabilistic (not deterministic) language	1
Inf-2	Properties of aggregates	-Being able to reason about a collection of data from individual cases as an aggregate	9
Inf-3	Sampling variability	<ul style="list-style-type: none"> <li>- Understanding the nature and behavior of sampling variability</li> <li>- Understanding sample to sample variability</li> <li>- Taking into account sample size in association with sampling variability</li> </ul>	2
Inf-4	The concept of unusualness	<ul style="list-style-type: none"> <li>-Being able to expect and reason whether or not a particular sample of data is likely given a particular expectation or claim (3)</li> <li>-Being able to describe the null model in the given context (4)</li> <li>-Being able to reason about unusualness of a sample statistic in the given context (5)</li> </ul>	3, 4, 5,
Inf-5	Relationship between sample size and distribution of sample statistics	-Being able to reason and articulate about the relationship between sample size and the shape of distribution of sample statistics	7
Inf-6	Generalizing from a sample to a population	<ul style="list-style-type: none"> <li>- Being able to draw a conclusion about a population from a sample based on the distribution of sample statistics (5)</li> <li>-Being able to make a conclusion about a population from a sample in association with change of sample size (8)</li> <li>- Being able to generalize (or make a conclusion) to a population using the null model and the distribution of sample statistics (recognizing the logic of statistical testing) (6)</li> </ul>	5, 6, 8

*(cont.)*

Topic Category	Topics	Learning Goals	Items
<i>Table D-1, cont.</i>			
Inf-7	Comparing two samples from two populations	<ul style="list-style-type: none"> <li>- Being able to predict and reason about possible differences between two populations based on observed differences between two samples of data (10, 11)</li> <li>- Being able to draw a conclusion about two populations (10)</li> <li>-Being able to take into account sample variations or sample size in relation with evidence to compare two samples (12, 13)</li> </ul>	10,11, 12, 13



Table D-2

*Test Blueprint to Assess Formal Inference*

Topic Category	Topics	Learning Goals	Items
Sampling distribution (SampD-1)	The concepts of samples and sampling	-Understanding the definition of sampling distribution -Understanding the role of sampling distribution	14
SampD-2	Sample representativeness	-Understanding importance of random sampling (recognizing biased sampling) (31) -Law of Large Numbers (Understanding that the larger the sample, the closer the distribution of the sample is expected to be to the population distribution)	31
SampD-3	Population distribution, sample distributions, and sampling distribution	-Understanding the relationship between sample distribution and population distribution (15) -Understanding the relationship between sampling distribution and population distribution (16)	15, 16
SampD-4	Central Limit Theorem	-Understanding the effect of sample size in sampling distributions (17) -Understanding how sampling error is related to making an inference about a sample mean	17
DE (DEsign of study)	Study design	-Understanding the logic of experimental design -Understanding difference between observational and experimental study -Understanding the purpose of random assignment in an experimental study	34
Statistical testing (Stest-1)	Definitions of $P$ -value and statistical significance	-Being able to recognize a correct interpretation of a $P$ -value (18) -Being able to calculate a numerical $P$ -value from a given distribution of statistics (25) -Being able to recognize a correct interpretation of statistical significance (27)	18, 25, 27
Stest-2	A statistical test for the comparison	-Being able to design a statistical test to compare two samples from two population (21, 22) -Designing a statistical test to compare two groups in an experiment -Being able to make a conclusion from a statistical test for comparing two groups	21, 22
Stest-3	Inference about a population proportion	-designing a statistical test for the proportion given in a sample (23) -making a conclusion about a statistical test for the population proportion (23)	23 <i>(cont.)</i>

Topic Category	Topics	Learning Goals	Items
<i>Table D-2, cont.</i>			
Stest-4	Inference about comparing two proportions	-being able to set up the null model to compare two proportions (24) -being able to make a conclusion about a statistical test for comparing two population proportions (26)	24, 26
CI (Confidence Interval)	Inference about Confidence Intervals	-Being able to interpret confidence interval in a given context (29) -Being able to interpret the relationship between confidence interval and margin of error (30)	29, 30
EV	Generalizing the results of ST Evaluation of ST	-Understanding that an experimental design with random assignment supports causal inference (20) -Understanding that an observational design with no random assignment doesn't support causal inference (28) -Being able to evaluate the results of hypothesis testing (considering sample size, practical significance, effect size, data quality, soundness of the method, etc.) (32, 33)	20, 28, 32, 33

## Appendix E

### Expert Review Forms of Preliminary Assessment

#### Item evaluation form (general)

##### Evaluation Form on the Assessment

This is an evaluation form to ask you to evaluate the assessment as a whole. The evaluation questions are intended to get information of how valid the proposed test is in assessing college students' informal and formal inference in statistics. If you haven't yet, please read each item and complete the evaluation question for each item before answering the items below.

Part 1. Please check the extent to which you agree or disagree with each of the following statements about the blueprint.

Item	Evaluation Questions	Ratings			
		Strongly agree ▼	Agree ▼	Disagree ▼	Strongly Disagree ▼
1	The items in the assessment are adequate to assess the learning goals specified in each category.				
2	The items in the assessment are related to the ISI.				
3	The items in the assessment are related to the FSI.				
4	The items in each category (ISI and FSI) are distinctive in terms of whether the item is categorized as one in ISI or FSI.				
5	The items are adequate to assess the construct of statistical inference.				

Part 2. For the following questions, please describe your opinions about the blueprint.

1. What do you think may be missing from the assessment items related to the constructs of informal and formal statistical inference?
2. What do you think of the assessment may be extraneous or not as important for assessing the constructs of informal and formal statistical inference?
3. Do you have any other suggestions for improving the assessment? Please describe.

Thank you!

### Item Evaluation Form (specific)

The following evaluation question was asked to the reviewers for each item (item 1-34).

Learning goal	<b>e.g.) Inf-1:</b> Being able to express uncertainty in making inference using probabilistic (not deterministic) language			
Please check the extent to which you agree or disagree with each of the following statements.	Ratings			
	Strongly Agree ▼	Agree ▼	Disagree ▼	Strongly Disagree ▼
This item assesses the stated learning goal.				
If you responded “Strongly disagree” or “Disagree”, please explain why you disagree and suggest how the item might be improved.				

## Appendix F

### Student Cognitive Interview Invitation

#### **Student Invitation Letter: Cognitive Interview**

To: Students who have taken EPSY 3264: Basic and Applied Statistics

You are invited to participate in a research study designed to develop and validate a research instrument called the *Assessment of Inferential Reasoning in Statistics (AIRS)*.

This instrument was developed to assess college students' statistical inference after they have taken an introductory statistics course. You were selected as a possible participant because you took an introductory statistics course last semester.

This study is being conducted by Jiyeon Park, a Ph.D student in the Department of Educational Psychology under the supervision of Dr. Robert delMas.

The study involves a one-hour interview where you will solve about 30 problems. You will be asked to talk aloud as you solve a set of 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 as you work through the assessment.

The problems may not look like anything you have done before and a problem may have several possible solutions 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 card.

The available times for the interview are:

Wednesday, July 13, 10am - 6pm

Thursday, July 14, 10am - 6pm

Friday, July 15, 2pm - 6pm

Monday, July 18 to Friday, July 22, 2pm - 6pm

If you are interested in participating please email me at [parkx666@umn.edu](mailto:parkx666@umn.edu) by this Friday, July 8. Please let me know all times that you are available on each day so that I can identify the best times for all students who want to participate.

You will be notified by Monday, July 11, if you are selected to participate in the study, and you will be told the time and location of the study at that time.

Thanks so much!

## **Consent Form: Student Cognitive Interview**

### **Consent Form: Think-alouds interview**

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 validate the "Assessment of Inferential Reasoning in Statistics (AIRS)". You were selected as a possible participant because you are currently taking or have taken post- secondary statistics courses. 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: Jiyoon Park, Educational Psychology, EPSY 5261 instructor

### **Background Information:**

The proposed study is to develop an instrument to assess two aspects of college students' statistical inferential reasoning—informal and formal statistical inference. The target population of the assessment is college students in the U.S. who are taking a non-calculus-based statistics course. The purposes of this assessment are: (1) to monitor students' longitudinal development of inferential reasoning as they learn statistics in an introductory course; and (2) to facilitate statistics education research on students' informal and formal statistical inference and the effect of instructional approaches on this topic.

### **Procedures:**

You will participate in a one-hour interview that is designed to gain an understanding of what reasoning and strategies you used for the questions in the AIRS 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 thinking and reasoning. 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 participation is the opportunity to develop a better understanding of statistics, and of your own statistical thinking.

**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. Research records will be kept in a locked file; 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 Jiyoong Park under the advisement of Professors Robert delMas, Ph.D. (Educational Psychology--Statistics Education) and Joan Garfield, Ph.D. (Educational Psychology—Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Jiyoong Park via my University of Minnesota, email: parkx666@umn.edu. You may also contact my advisor, Robert delMas, at delma001@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.

You need to 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 G

### Online Assessment Consent Form and Test Instruction

***Please read the description below and check in the Statement of Consent if you agree to participate in this study. \* This question is required***

You are invited to participate in a research study designed to develop and validate the *Assessment of Inferential Reasoning in Statistics (AIRS)*. You were selected as a possible participant because you are currently taking or have taken a post-secondary statistics course. Please read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by: Jiyoan Park, a Ph.D student in the department of Educational Psychology at the University of Minnesota.

#### **Background Information**

The purpose of this study is to develop an instrument to assess aspects of college students' statistical inferential reasoning. The target population of the assessment is students in the U.S. who are taking a non-calculus-based statistics course. The purposes of this assessment are: (1) to monitor the development of students' inferential reasoning as they learn statistics in an introductory course; and (2) to facilitate statistics education research on students' statistical inference and the effect of instructional approaches on this topic.

#### **Procedures**

If you agree to be in this study, you will take an online version of the assessment. The assessment consists of 34 questions and will take 40 to 50 minutes to complete.

#### **Risks and Benefits of Being in the Study**

There are no known risks to you as a participant. The benefit to participation is the opportunity to develop a better understanding of statistics, and of your own statistical thinking. 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. Any published report will not include any information that will make it possible to identify you as a participant. Research records will be kept in a locked file; 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 Jiyoan Park under the advisement of Professors Robert delMas, Ph.D. (Educational Psychology--Statistics Education) and Joan Garfield, Ph.D. (Educational Psychology—Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Jiyoan Park, at parkx666@umn.edu. You may also contact my advisor, Robert delMas, at delma001@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 check in the consent statement 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 part of this research project.

Please provide a unique code your instructor provided for your class. The code should be typed in capital letters (e.g., ABC or DEF01). \*

### **\*Online Test Instruction**

You will now start the AIRS online test. This test includes 34 multiple-choice type of questions. **Please read each question carefully and select the answer that best describes your reasoning.** You can click the **next button** to go the next question. You can also go back to previous question(s) to review or change your answer(s) by clicking the **back button**.

## Appendix H

### Expert Review on Test Blueprint

Table H-1

*Summary of Expert Comments*

	Comments and Suggestions	Who Commented	Change of the Current Blueprint	Rational for the Change
227	Common suggestions	In the category of Informal inference: There is no attention to inferences about the real world or contextual knowledge	Reviewer 1; Reviewer 2	Added some learning goals which consider <i>inferential reasoning in a given context</i>
		In categories of Formal inference (SD and ST): Too focus on the limited population	Reviewer 1: “one can conceptualize a process as an infinite, undefined population” Reviewer 3: “no comments are made about experiments”, only talk about samples from limited population.	Added the topics, DE (DEsign of study) and EV (evaluation of study) to get at students’ understanding of characteristics of different types of study in terms of— <i>how to design the study</i> and <i>how to generalize the results of the study</i>
		Need to have learning goals about understanding of effect size	Reviewer 2: In HT-1, Use the words “tool towards making a decision” Reviewer 3: For a HT showing a small P-value, we need to ask, “how large is the effect?” After that, we should consider data quality, soundness of the method etc.	In the category EV, added the learning goal, “Being able to evaluate the results of hypothesis testing considering —sample size, practical significance, effect size, data quality, soundness of the method, etc.

*(cont.)*

	Comments and Suggestions	Who Commented	Change of the Current Blueprint	Rational for the Change
	<i>Table H-1, cont.</i>			
Specific suggestions	Too focus on one type of problem, differences between groups, but almost half of the problems are about correlation problems (and regression)	Reviewer 1	Not included in the blueprint	Correlation and regression were considered as <i>literacy</i> or part of <i>descriptive statistics</i> rather than use of <i>inferential reasoning</i>
	Include learning goals about “Using models in informal inferential reasoning”		In two categories, informal inference and formal inference, the learning goals about setting up the null model in a given context was added.	
	Include using meta-cognitive awareness what inference is as opposed to performing some techniques		Not included in the blueprint	This learning goal was considered to be difficult to assess using typical test format (online format or paper-and-pencil format). Meta-cognitive awareness can be assessed through in-depth interview or individual observation.
				<i>(cont.)</i>

Comments and Suggestions	Who Commented	Change of the Current Blueprint	Rational for the Change
<i>Table H-1, cont.</i>			
Describe more explicitly about concepts like distribution, center and variation in aggregate category		In the category of <i>Properties of aggregates</i> the learning goal, <i>Being able to able to describe a collection of data using properties of distribution (shape, center, and variation but not necessarily using the terms)</i> , was added.	
Need to develop a topic category on Confidence Intervals	Reviewer 2	The topic category, “Inference about Confidence Interval, CI” was added.	
Need to consider data quality, soundness of the method etc.		The topic category, “Evaluation of HT (EV)”, was separated out from the Hypothesis Testing categories since this topic is more about assessing how to interpret and evaluate the results from statistical testing by integrating different kinds of information in a given study (e.g., random assignment, sample size, data quality). The learning goal about, “Being able to evaluate the results of hypothesis testing (considering sample size, practical significance, effect size, data quality, soundness of the method, etc.)”, was included in this EV category.	<i>(cont.)</i>

Comments and Suggestions	Who Commented	Change of the Current Blueprint	Rational for the Change
<i>Table H-1, cont.</i>			
In HT-6, add designing a test to compare two groups in an experiment. You might take samples from volunteers, not from populations.		In ST-3 (changed from category of HT), the learning goal, designing a statistical test to compare two groups in an experiment, was added.	
Consider including randomization and bootstrapping methods		Not included as a separate learning goals, but will be assessed in a way that items get at students reasoning of the ideas involved in randomization and bootstrap methods.  Considering that hypothesis testing based on normal distribution-based approach is not the only way of statistical testing, the original category about hypothesis testing (HT) was changed to statistical testing (ST), which includes randomization or bootstrap methods.	
For SD-2, in addition to “how larger samples look more like the population”, it is much more important “biased sampling” for sampling representativeness	Reviewer 3	The topic of “Law of Large Numbers” was changed to “sample representativeness” to assess whether students realize the importance of unbiased sampling (quality of samples) in addition to a large number of a sample (quantity of samples)	

Table H-2

*Detailed Comments*

Reviewer	Strongly disagree/Disagree to which evaluation question?	Why disagree? What suggestions to improve that part?	Any other suggestions?
Reviewer 1	<ul style="list-style-type: none"> <li>• Item 1. The topics of the blueprint represent the constructs of informal statistical inference</li> <li>• Item 3. The learning goals of the blueprint are adequate for developing items to asses students' understanding of informal statistical inference</li> <li>• Item 5. The set of learning goals is well supported by the literature</li> </ul>	<ul style="list-style-type: none"> <li>• There is no attention to inferences about the real world (contextual knowledge)</li> <li>• Limit focus to one type of problem, differences between groups, where almost half of the problems are about correlation problems (and regression)</li> <li>• using models in informal inferential reasoning</li> <li>• generalize to a process than to a population (one can conceptualize a process as an infinite, undefined population, but focus here is rather limited to finite population) – personally, processes are often more interesting than populations</li> </ul>	<ul style="list-style-type: none"> <li>• Add something like the role of inference in an investigative cycle, or in modeling.</li> <li>• Use of meta-cognitive awareness what inference is as opposed to performing some techniques</li> <li>• Including more explicitly concepts like distribution, center and variation in aggregate category</li> </ul>

*(cont.)*

Reviewer	Strongly disagree/Disagree to which evaluation question?	Why disagree? What suggestions to improve that part?	Any other suggestions?
<i>Table H-2, cont.</i>			
Reviewer 2	<ul style="list-style-type: none"> <li>Item 1. The topics of the blueprint represent the constructs of informal and formal inference.</li> <li>Item 2. The topics of the blueprint represent the constructs of formal statistical inference</li> <li>Item 4. The learning goals of the blueprint are adequate for developing items to assess students' understanding of formal statistical inference</li> <li>Item 8. The blueprint provides a framework of developing a test to assess informal and formal statistical inference</li> </ul>	<ul style="list-style-type: none"> <li><u>For informal inference:</u> <ul style="list-style-type: none"> <li>- "Inf-5: Generalizing from a sample to population", consider use of "contextual knowledge". Can ask, "Can the conclusion make sense?" or "Alternative factors or explanations?"</li> <li>- students' realizing the link between sample and population</li> </ul> </li> <li>Reasoning about comparison of two groups in an experiment.</li> <li>Student misconceptions about the relationship between sample distribution, sampling distribution, and population distribution</li> <li><u>For Hypothesis testing:</u> <ul style="list-style-type: none"> <li>very focused on the P-value. Need to develop a topic category on Confidence Intervals.</li> <li>In HT-1. Use the words "tool towards making a decision". For a HT showing a small P-value, we need to ask, "how large is the effect?". After that, we should consider data quality, soundness of the method etc.</li> <li>In HT-6, change the sentence to comparing two populations based on a sample from each population</li> <li>In HT-6, add designing a test to compare two groups in an experiment. You might take samples from volunteers, not from populations.</li> </ul> </li> <li><u>For formal inference:</u> <ul style="list-style-type: none"> <li>Consider including randomization and bootstrapping methods: the current blueprint assumes that norm-based inference is the only method for inference yet statistical practice is very quickly adopting these methods.</li> </ul> </li> </ul>	

*(cont.)*

Reviewer	Strongly disagree/Disagree to which evaluation question?	Why disagree? What suggestions to improve that part?	Any other suggestions?
<i>Table H-2, cont.</i>			
Reviewer 3	He “strongly agreed” or “agreed” for every evaluation question.	<ul style="list-style-type: none"> <li>• <u>For informal inference:</u> <ul style="list-style-type: none"> <li>-Inf-5 and Inf-6 both talk about generalizing to a population, but no comments are made about experiments.</li> <li>-In Inf-3, inference about effect size and data variability need to be included.</li> </ul> </li> <li>• <u>For formal inference:</u> <ul style="list-style-type: none"> <li>-For SD-2, in addition to “how larger samples look more like the population”, it is much more important “biased sampling” for sampling representativeness.</li> <li>-Like in Informal inference, effect size and data variability are important topics.</li> </ul> </li> </ul>	



Appendix I  
 Versions of Assessment  
**Preliminary Version**

*Assessment of Inferential Reasoning in Statistics (AIRS)*

*[NOTE: The free-response format will be revised to multiple-choice format after piloting.]*

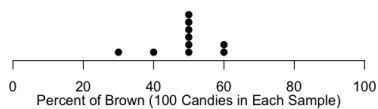
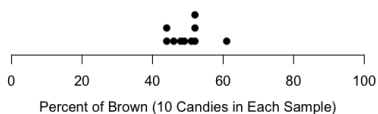
**Informal inferential reasoning items**

1. The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched their records for those days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days. The forecast of 70% chance of rain can be considered very accurate if it rained on:

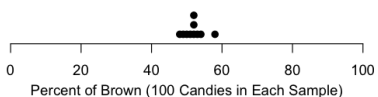
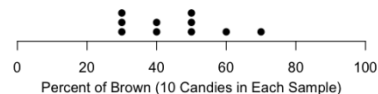
- a. 95% - 100% of those days.
- b. 85% - 94% of those days.
- c. 75% - 84% of those days.
- D. 65% - 74% of those days.**
- e. 55% - 64% of those days.

2. Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 50% brown candies. Ten students each take one random sample of 10 candies and record the percentage of brown candies in each of their samples. Another ten students each take one random sample of 100 candies and record the percentage of brown candies in each of their samples. Which of the following pairs of graphs represent the most plausible distributions for the percent of brown candies obtained in the samples for each group of 10 students?

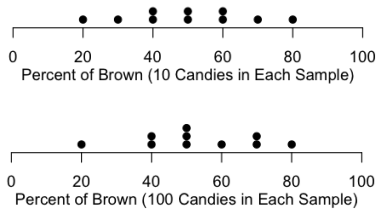
a.



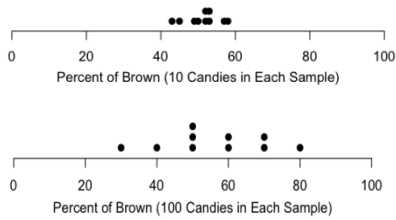
**B.**



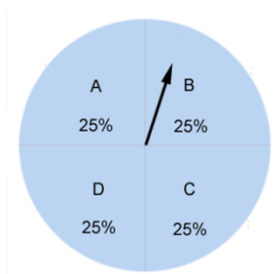
c.



d.



**Question 3 to 9 refer to the following:** Consider a spinner shown below that has the letters from *A* to *D*.



Let's say you used the spinner 10 times and each time you wrote down the letter that the spinner lands on. Furthermore, let's say when you looked at the results, you saw that the letter *B* showed up 5 times out of the 10 spins.

Suppose a person is watching you play the game and they say that it seems like you got too many *B*'s.

A second person says that 5 *B*'s would not be unusual for this spinner.

3. If the spinner is fair, how many *B*'s out of 10 spins would you expect to see?

- A. 2 or 3 *B*'s
- b. 4 or 5 *B*'s
- c. 6 or 7 *B*'s
- d. 8 or 9 *B*'s

4. Which person do you think is correct? And why?

a. The first person because:.

**B.** The second person because:

c. Both are correct because:

5. A statistician wants to set up a probability model to examine how often the result of 5 B's out of 10 spins could happen with the spinner just by chance alone. What would be the probability model the statistician can use to do a test? Please describe the null model.

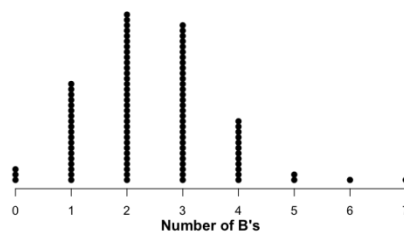
a. All the trials of getting letters are independent.

B. The probability for each letter is  $p(A)=1/4$ ,  $p(B)=1/4$ ,  $p(C)=1/4$ ,  $p(D)=1/4$ .

c. The probability for letter B is  $1/2$  and the other three letters each have probability of  $1/6$ .

d. The probability for letter B is  $1/2$  and the probabilities for the other letters sum to  $1/2$ .

6. The following dot plot represents the distribution for the number of B's that the statistician got based on the null model from 100 samples where each sample consisted of the results from 10 spins. What do you think about the observed result of 5 B's? [\*Free-response question]



a. 5 B's are not unusual because:

b. 5 B's are unusual because:

c. There is not enough information to decide if 5 B's is unusual or not.

7. Based on your answers to the questions 4 and 5, what would you conclude about whether or not the spinner is fair? Explain your reasoning. [\*Free-response question]

a. This spinner is fair because:

b. This spinner is unfair because:

\*Note: This item will be revised to multiple-choice format after piloting based on student responses.

8. Let's say you try the spinner again to gather more data. You spin it 20 times and get the same *proportion* of B's as before, (10 B's out of the 20 times, or  $\frac{1}{2}$  B's). How would you expect the distribution of the *proportion* of B's obtained from 100 samples of 20 spins each to compare to the distribution of the *proportion* of B's obtained from 100 samples of 10 spins each?

a. The distribution of the proportion of B's for 100 samples of 20 spins each would be wider because you have twice as many spins in each trial.

**B.** The distribution of the proportion of B's for 100 repetitions of 20 spins each would be narrower because you have more information for each sample.

c. Both distributions would have about the same width because the probability of getting each letter is the same whether you do 10 spins or 20 spins.

9. Which situation, 5 B's out of 10 spins or 10 B's out of 20 spins, provides the stronger evidence that the spinner is not fair? Explain your reasoning. [\*Free-response question]

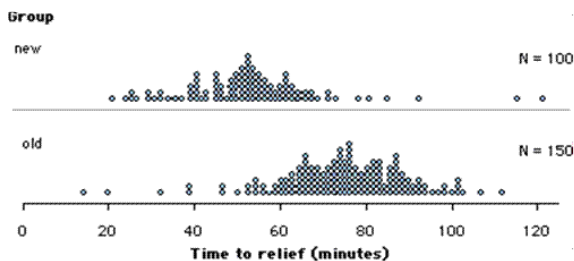
A. 10 B's out of 20 spins because:

b. 5 B's out of 10 spins because:

c. Both outcomes provide the same evidence because:

\*Note: This item will be revised to multiple-choice format after piloting based on student responses.

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



a. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.

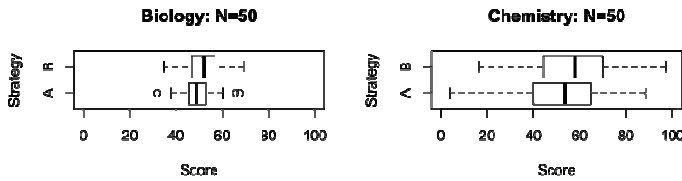
b. The average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.

c. We can't conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.

**Question 11 and 12 refer to the following:** An experiment was designed to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects are randomly assigned to each exam preparation strategy. After completing the exam preparation, all subjects take the same exam (which is scored from 0 to 100). Four different experiments are conducted with students who are enrolled in introductory courses for four different subject areas: (biology, chemistry, psychology, sociology)

The dot plots in question 10 and 11 are distributions of exam scores obtained from two experiments, where the subjects prepared with two different strategies, A and B.

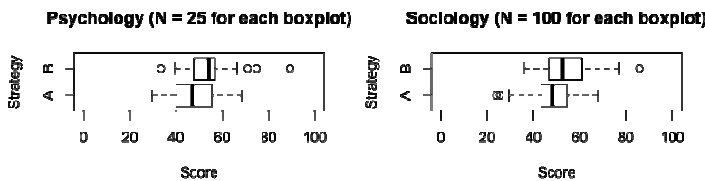
11. Boxplots of exam scores for students in the biology course are shown below on the left, and the boxplots for the students in the chemistry course are on the right. For each subject area, 25 students were randomly assigned to either strategy A and 25 students were randomly assigned to strategy B. Which subject area, biology or chemistry, provides the stronger evidence against the claim, “neither strategy is better than the other”? Select either Biology or Chemistry and right an explanation for your choice.



- A. Biology
  - b. Chemistry
- Explain your choice:

\*Note: This item will be revised to multiple-choice format after piloting based on student responses.

12. Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students who were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to either strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence against the claim, “neither strategy is better than the other”? Why?



- a. Psychology
  - B. Sociology**
- Explain your choice:

\*Note: This item will be revised to multiple-choice format after piloting based on student responses.

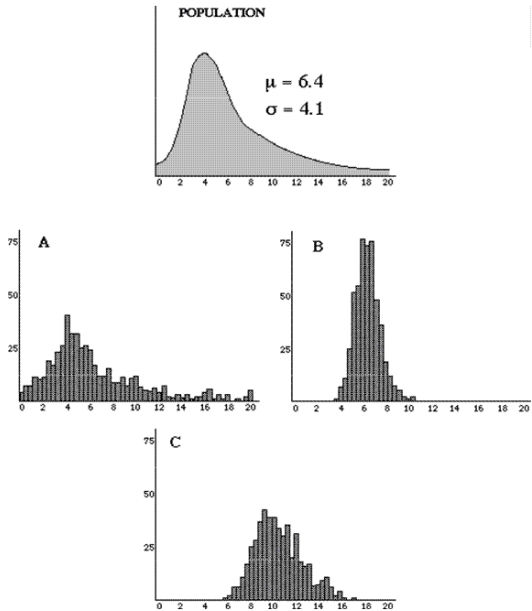
**Formal inferential reasoning items**

- 13. A random sample of 25 textbooks for different courses taught at a University is obtained, and the mean textbook price is computed for the sample. To determine the probability of finding another random sample of 25 textbooks with a mean more extreme than the one obtained from this random sample, you would need to refer to:
  - a. the distribution of textbook prices for all courses at the University.

- b. the distribution of textbook prices for this sample of University textbooks.
- C.** the distribution of mean textbook prices for all samples from the University.

14 – 15. Items 14 and 15 refer to the following situation:

Four graphs are presented below. The graph at the top is a distribution for a population of test scores. The mean score is 6.4 and the standard deviation is 4.1.



14. Which graph (A, B, or C) do you think represents a single random sample of 500 values from this population?
- A. Graph A
  - b. Graph B
  - c. Graph C
15. Which graph (A, B, or C) do you think represents a distribution of 500 sample means from random samples each of size 9?
- a. Graph A
  - B.** Graph B
  - c. Graph C
16. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?
- a. A random sample of a sample size of 100 with a sample mean of 12.1.
  - b. A random sample of a sample size of 36 with a sample mean of 11.5.
  - C.** A random sample of a sample size of 100 with a sample mean of 11.5
  - d. A random sample of a sample size of 36 with a sample mean of 12.1

17. A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes ( $p = 0.287$ ). This means

- a. The distribution of the GPAs for male and female scholarship athletes are identical except for 28.7% of the athletes.
- b. The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287.
- c. There is a 0.287 chance that a pair of randomly chosen male and female scholarship athletes would have a significant difference.

**D.** There is a 0.287 chance of obtaining as large or larger of a mean difference in GPAs between male and female scholarship athletes as that observed in the sample.

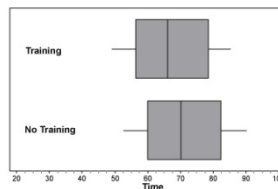
Questions 18 and 19 refer to the following: A researcher investigates the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either be exposed or not be exposed to the herbicide. The fish exposed to the herbicide showed higher levels of an enzyme associated with cancer.

- 18. Suppose no statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?
  - a. The researcher must not be interpreting the results correctly; there should be a significant difference.
  - b. The sample size may be too small to detect a statistically significant difference.
  - c. It must be true that the herbicide does not cause higher levels of the enzyme.
- 19. Suppose a statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?
  - a. There is evidence of association, but no causal effect of herbicide on enzyme levels.
  - b. The sample size is too small to draw a valid conclusion.
  - c. He has proven that the herbicide causes higher levels of the enzyme.
  - d. There is evidence that the herbicide causes higher levels of the enzyme for these fish.

**20 – 21. Read the following information to answer questions 20 and 21:**

Data are collected from a research study that compares performance for professionals who have participated in a new training program with performance for professionals who haven't participated in the program. The professionals are randomly assigned to one of two groups, with one group being given the new training program and the other group being not given. For each of the following pairs of graphs, indicate what you would do next to determine if there is a statistically significant difference between the training and no training groups.

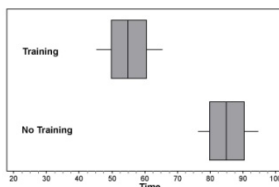
**20.**



- a. Nothing, the two groups appear to be statistically significantly different.

b. Conduct an appropriate statistical test for a difference between groups

21.



A. Nothing, the two groups appear to be statistically significantly different

B. Conduct an appropriate statistical test for a difference between groups

**Read the following information to answer Question 22:**

A student participates in a Coke versus Pepsi taste test. She correctly identifies which soda is which seven times out of ten tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You want to estimate the probability that this student could get *at least seven right out of ten tries just by chance alone*.

You decide to follow a procedure where you:

- Simulate a chance process in which you specify the **probability** of making a correct guess on each trial
- Repeatedly generate ten cases per trial from this process and record the number of correct outcomes in each trial
- Calculate the proportion of trials where the number of correct guesses meets a **specified criterion**

In order to run the procedure, you need to decide on the value for the probability of making a correct guess, and specify the criterion for the number of correct guesses.

**22. Which of the options below would provide a reasonable approach to simulating data in order to determine the probability of anyone getting seven out of ten tries correct just by chance alone?**

- Specify the probability of a correct guess as 50% and calculate the proportion of all trials with *exactly seven* correct guesses
- Specify the probability of a correct guess as 50% and calculate the proportion of all trials with *seven or more* correct guesses**
- Specify the probability of a correct guess as 70% and calculate the proportion of all trials with *exactly seven* correct guesses
- Specify the probability of a correct guess as 70% and calculate the proportion of all trials with *seven or more* correct guesses

**Read the following information before answering Questions 23– 25:**

A research question of interest is whether financial incentives can improve performance. Alicia designed a study to test whether video game players are more likely to win on a certain video game when offered a \$5 incentive compared to when simply told to “do your best.” Forty subjects are randomly assigned to one of



two groups, with one group being offered \$5 for a win and the other group simply being told to “do your best.” She collected the following data from her study:

	\$5 incentive	“Do your best”	Total
Win	16	8	24
Lose	4	12	16
Total	20	20	40

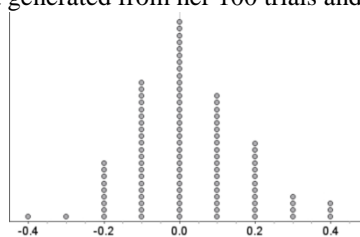
It looks like the \$5 incentive is more successful than the encouragement. The difference in success rates as a proportion is

$$\frac{16}{20} - \frac{8}{20} = 0.40$$

In order to test whether this apparent difference might be due simply to chance, she does the following:

- She gets 40 index cards. On 24 of the cards she writes "win" and on 16 she writes "lose".
  - She then shuffles the cards and randomly places the cards into two stacks. One stack represents "\$5 incentive" and the other "verbal encouragement".
  - For this simulation, she computes the observed difference in the success rates by subtracting the success rate for the simulation's "\$5 incentive" group from the success rate of the simulation's "verbal encouragement" group.
- She repeats the previous two steps 100 times.
- She plots the 100 statistics she observes from these trials.

This is the simulated data that Alicia generated from her 100 trials and used to test her research question:



**23. What is the null model that Alicia's data simulated?**

- a. The \$5 incentive is more effective than verbal encouragement for improving performance.
- b. The \$5 incentive and verbal encouragement are equally effective at improving performance.**
- c. Verbal encouragement is more effective than a \$5 incentive for improving performance.

**24. Use this distribution to estimate the  $p$ -value for her observed result. Explain how you got the  $p$ -value.**

- a. 0.02
- b. 0.03**
- c. 0.04**
- d. 0.05

- e. 0.4
- f. 0.5

**Explain your choice:**

**25. What does the distribution tell you about the hypothesis that \$5 incentives are effective for improving performance?**

- a. The incentive is not effective because the null distribution is centered at 0.
- b. The incentive is effective because the null distribution is centered at 0.
- c. The incentive is not effective because the  $p$ -value is greater than .05
- d. The incentive is effective because the  $p$ -value is less than .05**

**Questions 26 to 29 refer to the following:** Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took a college admissions test twice. Of these, 500 had taken coaching courses between their two attempts at the college admissions test. The study compared the average increase in scores (out of the total possible score of 2,400) for students who were coached with the average increase for students who were not coached.

26. The result of this study showed that when students retake the SAT test, the difference between the average increase for coached and not-coached students was not statistically significant. This means that
- a. The sample sizes were too small to detect a true difference between the coached and not-coached students.
  - b. The difference between coached and not-coached students could occur just by chance even if coaching really has no effect.
  - c. The increase in test scores makes no difference in getting into college since it is not statistically significant.
  - d. The study was badly designed because they did not have equal numbers of coached and not-coached students.
27. The study doesn't show that coaching causes a greater increase in SAT scores. One plausible reason is that
- a. the not-coached students used other effective ways to prepare.
  - b. 4,200 students is too few to draw a conclusion.
  - c. more students were not coached than were coached.
  - d. Students were not randomly assigned to the two groups.
28. The report of the study states, "With 95% confidence, we can say that the average score for students who take the college admissions test a second time is between 28 and 57 points higher than the average score for the first time." By "95% confidence" we mean:
- a. 95% of all students will increase their score by between 28 and 57 points for a second test.
    - b. We are certain that the average increase is between 28 and 57 points.
    - c. We got the 28 to 57 point higher mean scores in a second test in 95% of all samples.
    - d. 95% of all adults would believe the statement.

29. We are 95% confidence that the difference between average scores for coached and uncoached students is between 28 and 57 points. If we want to be 99% confident, the range of points would be:
- Wider, because higher confidence requires a larger margin of error.
  - Narrower, because higher confidence requires a smaller margin of error.
  - Exactly the same width as for 95% confidence.

**Questions 30 to 31 refer to the following:** Sale of eggs that are contaminated with salmonella can cause food poisoning among consumers. A large egg producer takes a random sample of 200 eggs from all the eggs shipped in one day. The laboratory reports that 9 of these eggs had salmonella contamination. Unknown to the producer, 0.1% (one-tenth of one percent) of all eggs shipped had salmonella.

30. A statistician tells the producer that the margin of error for a 95% confidence statement for these data is about plus or minus 3percentage points. The producer therefore reports that between 1.5% and 7.5% (that's 4.5%  $\pm$  3%) of all eggs are contaminated. This isn't right because only 0.1% of all eggs from the producer are contaminated. What went wrong?
- The statement that 0.1% of all of the eggs shipped were contaminated with salmonella must be wrong; it has to be at least 1.5% of all eggs shipped.
  - A 95% confidence statement is only right for 95% of all possible samples. This must be one of the 5% of samples for which we get an incorrect conclusion.
  - The laboratory tests must be wrong because it's impossible for the true percentage to lie outside the confidence interval.
31. If the producer took an random sample of 400 eggs instead of 200, the new margin of error would be:
- The same as before, because the population of eggs is the same.
  - Smaller than before, because the sample is larger.
  - Larger than before, because the sample is larger.
  - Random in size, could be either larger or smaller than before.
  - Can't tell, because sample size doesn't control the margin of error.
32. A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?
- This is a simple random sample. It will give an accurate estimate.
  - Because the sample is so small, it will not give an accurate estimate.
  - This is a census, because all fans had a chance to be asked. It will give an accurate estimate.
  - The sampling method is biased. It will not give an accurate estimate.**
33. Suppose we wish to estimate the percentage of students who smoke cigarettes at each of several colleges and universities. One is a small liberal arts college with an enrollment 2,000 undergraduates and another is a large public university with an enrollment of 30,000 undergraduates. A simple random sample of 5% of the students is taken at each school and used to estimate the percentage of students who smoke. The margin of error for the estimate will be:
- smaller for the liberal arts college.
  - smaller for the university.

- d. about the same at both schools.
- e. anything - you can't tell without seeing the sample results.

34. A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that prescription medications only were the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?
- a. Yes, because medication patients lived longer.
  - b. No, because doctors chose the treatments.**
  - c. Yes, because the study was a comparative experiment.
  - d. No, because the patients volunteered to be studied.
35. An engineer designs an improved light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of 40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. The most likely explanation is
- a. The new design had more variability than the previous design.
  - b. The sample size for the new design is very large.
  - c. The mean of 1,200 for the previous design is large.
36. Research participants were randomly assigned to take Vitamin E or a placebo pill. After taking the pills for eight years, it was reported how many developed cancer. Which of the following responses gives the best explanation as to the purpose of randomization in this study?
- a. To ensure that all potential cancer patients had an equal chance of being selected for the study.
  - b. To reduce the amount of sampling error.
  - c. To produce treatment groups with similar characteristics.**
  - d. To prevent skewness in the results.

===The End ===

## AIRS-1 (Changes were made from expert reviews)

### *Assessment of Inferential Reasoning in Statistics-1 (AIRS-1)*

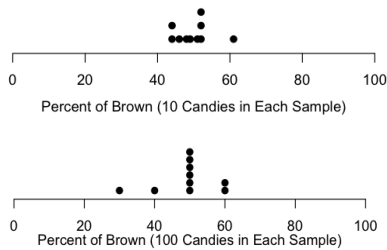
**[NOTE: The free-response format will be revised to multiple-choice format after piloting.]**

1. The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched their records for those days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days. The forecast of 70% chance of rain can be considered very accurate if it rained on:

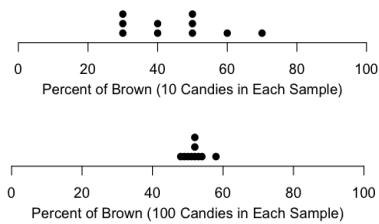
- a. 95% - 100% of those days.
- b. 85% - 94% of those days.
- c. 75% - 84% of those days.
- d. 65% - 74% of those days.
- e. 55% - 64% of those days.

2. Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 50% brown candies. Ten students each take one random sample of 10 candies and record the percentage of brown candies in each of their samples. Another ten students each take one random sample of 100 candies and record the percentage of brown candies in each of their samples. Which of the following pairs of graphs represents the more plausible distributions for the percentage of brown candies obtained in the samples for each group of 10 students?

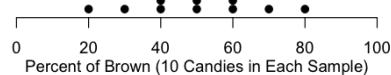
a.

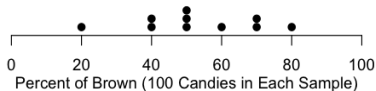


b.

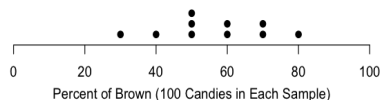
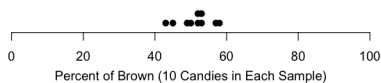


c.

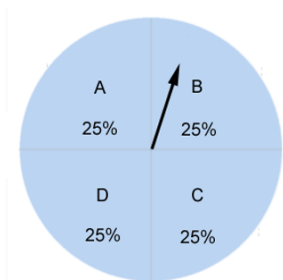




d.



**Question 3 to 9 refer to the following:** Consider a spinner shown below that has the letters from *A* to *D*.



'Person 1' used the spinner 10 times and each time he wrote down the letter that the spinner landed on. When he looked at the results, he saw that the letter *B* showed up 5 times out of the 10 spins. Now he doubts the fairness of the spinner because it seems like he got too many *B*s. However, 'Person 2' says that 5 *B*s would not be unusual for this spinner.

4. If the spinner is fair, how many *B*s out of 10 spins would you expect to see?
  - a. 2 or 3 *B*'s
  - b. 4 or 5 *B*'s
  - c. 6 or 7 *B*'s
  - d. 8 or 9 *B*'s

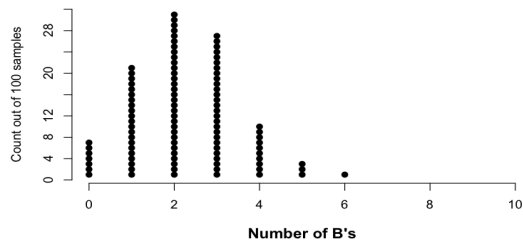
4. Which person do you think is correct and why?

- a. Person 1 is correct because:
- b. Person 2 is correct because:
- c. Both are correct because:

5. A statistician wants to set up a probability model to examine how often the result of 5 *B*'s out of 10 spins could happen with a fair spinner just by chance alone. Please describe the null model. [\*Free-response question]

6. The statistician conducted a statistical test to examine the fairness of the spinner using a computer simulation. The computer simulation randomly generates four letters, *A* to *D*. She obtained 100 samples

where each sample consisted of 10 letters. She then counted the number Bs in each sample of 10 random letters. The following dot plot represents the number of Bs for each of the 100 samples. What do you think about the observed result of 5 Bs out of 10 spins in the spinner?



- 5 B's are not unusual because:.
- 5 B's are unusual because:.
- There is not enough information to decide if 5 B's is unusual or not.

7. Based on your answers to questions 5 and 6, what would you conclude about whether or not the spinner is fair? Why? [\*Free-response question]

- This spinner is fair because:
- This spinner is unfair because:

\*Note: This item will be revised to multiple-choice format after piloting based on student responses.

8. Let's say the statistician did another computer simulation, but this time each sample consisted of 20 spins. She calculated the proportion of Bs in each sample (the number of Bs divide by 20). How would you expect the distribution of the *proportion* of Bs obtained from 100 samples of 20 spins each to compare to the distribution of the *proportion* of Bs obtained from 100 samples of 10 spins each?

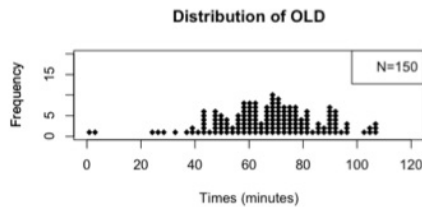
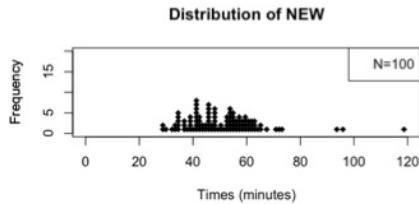
- The distribution of the proportion of Bs for 100 samples of 20 spins each would be wider because you have twice as many spins in each trial.
- The distribution of the proportion of Bs for 100 repetitions of 20 spins each would be narrower because you have more information for each sample.
- Both distributions would have about the same width because the probability of getting each letter is the same whether you do 10 spins or 20 spins.

9. Which of the following results, 5 Bs out of 10 spins or 10 Bs out of 20 spins, provides the stronger evidence that the spinner is not fair? Explain your reasoning.

- 10 Bs out of 20 spins because larger samples have less variability, so it is less likely to get an unusual result with a fair spinner.
- 5 Bs out of 10 spins because smaller samples have larger variability, so it is more likely to get an unusual result with a fair spinner.
- Both outcomes provide the same evidence because there is the same proportion of Bs ( $1/2$ ) in each of the two samples.

**Item 10 to 12 refers to the following situation:**

A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown below. Questions 9, 10, and 11 present statements made by three different statistics students. For each statement, indicate whether you think the student's conclusion is valid.



10. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.

- Valid
- Not valid

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

- Valid
- Not valid

12. We can't conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.

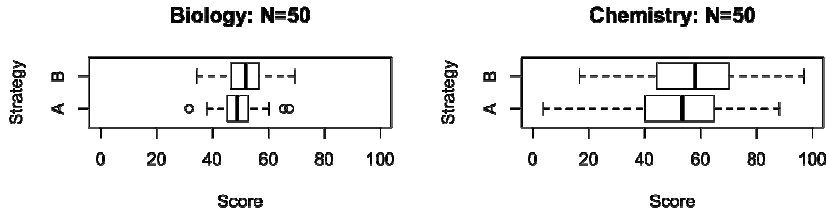
- Valid
- Not valid

**Question 13 and 14 refer to the following:** Four experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100) in all four experiments. The four different experiments were conducted with students who were enrolled in four different subject areas: biology, chemistry, psychology, sociology.

13. Boxplots of exam scores for students in the biology course are shown below on the left, and the

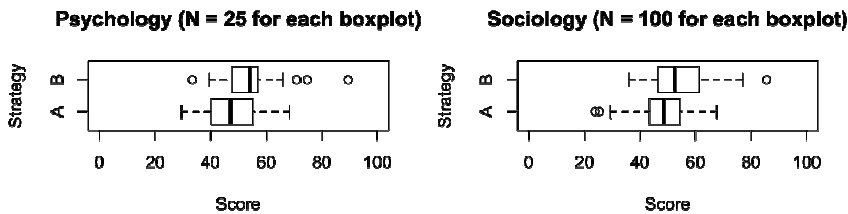


boxplots for the students in the chemistry course are on the right. For each subject area, 25 students were randomly assigned to either strategy A and 25 students were randomly assigned to strategy B. Which experiment, the one for the biology or the chemistry course, provides the stronger evidence against the claim, “neither strategy is better than the other”? Why?



- a. Biology
- b. Chemistry

14. Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence against the claim, “neither strategy is better than the other”? Why?



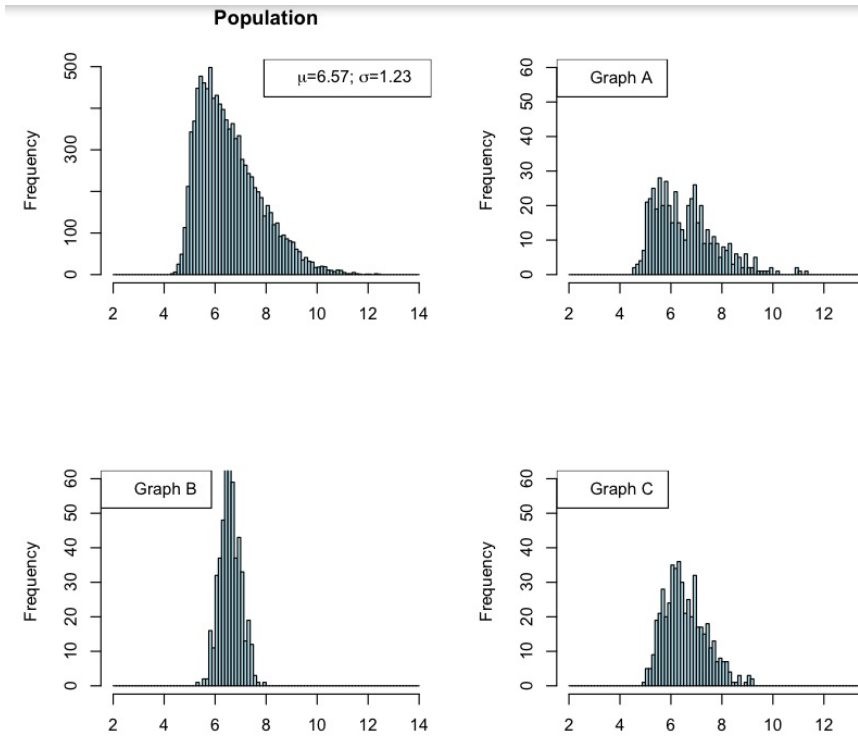
- a. Psychology
- b. Sociology

15. A random sample of 25 textbooks for different courses taught at a University is obtained, and the mean textbook price is computed for the sample. To determine the probability of finding another random sample of 25 textbooks with a mean more extreme than the one obtained from this random sample, you would need to refer to:

- a. the distribution of textbook prices for all courses at the University.
- b. the distribution of textbook prices for this sample of University textbooks.
- c. the distribution of mean textbook prices for all samples of size 25 from the University.

Questions 16 and 17 refer to the following situation:

Four graphs are presented below. The graph at the top is a distribution for a population of test scores. The mean score is 6.57 and the standard deviation is 1.23.



16. Which graph (A, B, or C) do you think represents a single random sample of 500 values from this population?
- Graph A
  - Graph B
  - Graph C
17. Which graph (A, B, or C) do you think represents a distribution of 500 sample means from random samples each of size 9?
- Graph A
  - Graph B
  - Graph C
18. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?
- A random sample of a sample size of 100 with a sample mean of 12.1.
  - A random sample of a sample size of 36 with a sample mean of 11.5.

- c. A random sample of a sample size of 100 with a sample mean of 11.5.
- d. A random sample of a sample size of 36 with a sample mean of 12.1.

19. A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes ( $p = 0.287$ ). What does this mean?

- a. The distribution of the GPAs for male and female scholarship athletes are identical except for 28.7% of the athletes.
- b. The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287.
- c. There is a 28.7% chance that a pair of randomly chosen male and female scholarship athletes would have a significant difference assuming that there is no difference.
- d. There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between male and female scholarship athletes as that observed in the sample assuming that there is no difference.

Questions 20 and 21 refer to the following: A researcher investigates the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either be exposed or not be exposed to the herbicide. The fish exposed to the herbicide showed higher levels of an enzyme associated with cancer.

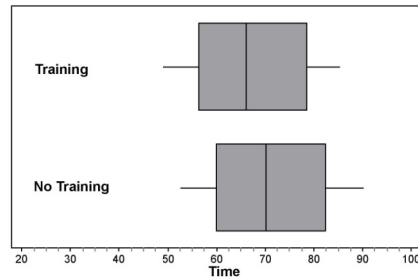
20. Suppose no statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?
- a. The researcher must not be interpreting the results correctly; there should be a significant difference.
  - b. The sample size may be too small to detect a statistically significant difference.
  - c. It must be true that the herbicide does not cause higher levels of the enzyme.
21. Suppose a statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?
- a. There is evidence of association, but no causal effect of herbicide on enzyme levels.
  - b. The sample size is too small to draw a valid conclusion.
  - c. He has proven that the herbicide causes higher levels of the enzyme.
  - d. There is evidence that the herbicide causes higher levels of the enzyme for these fish.

**22 – 23. Read the following information to answer questions 20 and 21:**

Data are collected from a research study that compares the times to complete a task for professionals who have participated in a new training program with performance for professionals who haven't participated in the program. The professionals are randomly assigned to one of the two groups, with one group receiving the new training program ( $N=50$ ) and the other group not receiving the training ( $N=50$ ). For each of the

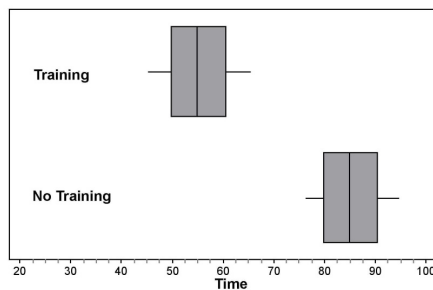
following pairs of graphs, select an appropriate action that you would need to do next to determine if there is a statistically significant difference between the training and no training groups. Write an explanation for your choice.

22.



- Nothing, the two groups appear to be statistically significantly different.
- Conduct an appropriate statistical test for a difference between groups.

23.



- Nothing, the two groups appear to be statistically significantly different.
- Conduct an appropriate statistical test for a difference between groups.

24. A student participates in a Coke versus Pepsi taste test. She identifies the correct soda seven times out of ten tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You want to estimate the probability that this student could get *at least seven right out of ten tries just by chance alone*.

You decide to follow a procedure where you:

- Simulate a chance process in which you specify the **probability** of making a correct guess on each trial
- Repeatedly generate ten cases per trial from this process and record the number of correct outcomes in each trial
- Calculate the proportion of trials where the number of correct guesses meets a **specified criterion**

In order to run the procedure, you need to decide on the value for the probability of making a correct guess, and specify the criterion for the number of correct guesses.

Which of the options below would provide a reasonable approach to simulating data in order to determine the probability of anyone getting seven out of ten tries correct just by chance alone?

- Specify the probability of a correct guess as 50% and calculate the proportion of all trials with *exactly seven* correct guesses
- Specify the probability of a correct guess as 50% and calculate the proportion of all trials with *seven or more* correct guesses
- Specify the probability of a correct guess as 70% and calculate the proportion of all trials with *exactly seven* correct guesses
- Specify the probability of a correct guess as 70% and calculate the proportion of all trials with *seven or more* correct guesses

**Read the following information before answering Questions 25– 26:**

A research question of interest is whether financial incentives can improve performance. Alicia designed a study to test whether video game players are more likely to win on a certain video game when offered a \$5 incentive compared to when simply told to “do your best.” Forty subjects are randomly assigned to one of two groups, with one group being offered \$5 for a win and the other group simply being told to “do your best.” She collected the following data from her study:

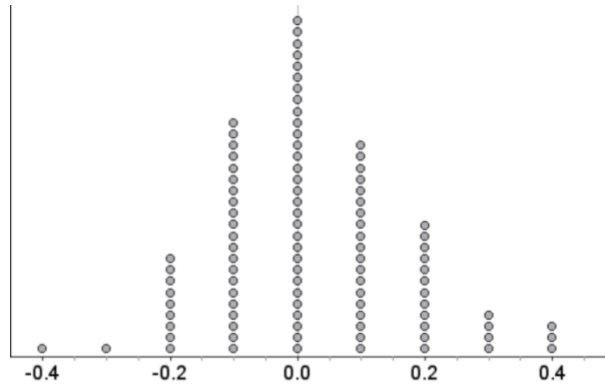
	\$5 incentive	“Do your best”	Total
Win	16	8	24
Lose	4	12	16
Total	20	20	40

It looks like the \$5 incentive is more successful than the encouragement. The difference in success rates as a proportion is:  $16/20 - 8/20 = 8/20 = 0.40$

In order to test whether this apparent difference might be due simply to chance, she does the following:

- She gets 40 index cards. On 24 of the cards she writes "win" and on 16 she writes "lose".
  - She then shuffles the cards and randomly places the cards into two stacks. One stack represents "\$5 incentive" and the other "verbal encouragement".
  - For this simulation, she computes the observed difference in the success rates by subtracting the success rate for the simulation's "\$5 incentive" group from the success rate of the simulation's "verbal encouragement" group.
- She repeats the previous two steps 100 times.
- She plots the 100 statistics she observes from these trials.

The following shows a distribution of simulated data that Alicia generated from her 100 trials and used to test her research question:



25. What is the null model (null hypothesis) that Alicia's data simulated?
- The \$5 incentive is more effective than verbal encouragement for improving performance.
  - The \$5 incentive and verbal encouragement are equally effective for improving performance.
  - Verbal encouragement is more effective than a \$5 incentive for improving performance.
26. Use this distribution to estimate the  $p$ -value for her observed result.
- 0.02
  - 0.03
  - 0.04
  - 0.05
  - 0.40
27. What does the distribution tell you about the hypothesis that \$5 incentives are effective for improving performance?
- The incentive is not effective because the null distribution is centered at 0.
  - The incentive is effective because the null distribution is centered at 0.
  - The incentive is not effective because the  $p$ -value is greater than .05.
  - The incentive is effective because the  $p$ -value is less than .05.

**Questions 28 to 31 refer to the following:** Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached.

28. The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean?
- The sample sizes were too small to detect a true difference between the coached and not-coached students.
  - The observed difference between coached and not-coached students could occur just by chance alone even if coaching really has no effect.

- c. The increase in test scores makes no difference in getting into college since it is not statistically significant.
  - d. The study was badly designed because they did not have equal numbers of coached and not-coached students.
29. The study doesn't show that coaching causes a greater increase in college admissions test scores. Which of the following would be the most plausible reason for this?
- a. The not-coached students used other effective ways to prepare.
  - b. The number of 4,200 students is too few to detect a difference.
  - c. More students were not coached than were coached.
30. The report of the study states, "With 95% confidence, we can say that the average score for students who take the college admissions test a second time is between 28 and 57 points higher than the average score for the first time." By "95% confidence" we mean:
- a. 95% of all students will increase their score by between 28 and 57 points for a second test.
  - b. 95% of all samples of students will increase their score by between 28 to 57 points for a second test.
  - c. 95% of all students who take the college admissions test would believe the statement.
  - d. We are 95% certain that the average increase in college admissions scores is between 28 and 57 points.
31. We are 95% confident that the difference between average scores for the first and the second tests is between 28 and 57 points. If we want to be 99% confident, the range of values in the interval would be:
- a. Wider, because higher confidence requires a larger margin of error.
  - b. Narrower, because higher confidence requires a smaller margin of error.
  - c. Exactly the same width as the range for the 95% confidence interval.
32. A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?
- a. This is a simple random sample. It will give an accurate estimate
  - b. Because the sample is so small, it will not give an accurate estimate
  - c. Because all fans had a chance to be asked, it will give an accurate estimate.
  - d. The sampling method is biased. It will not give an accurate estimate.
33. A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that 'prescription medications only' was the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?

- a. Yes, because medication patients lived longer.
- b. No, because doctors chose the treatments.
- c. Yes, because the study was a comparative experiment.
- d. No, because the patients volunteered to be studied.

34. An engineer designs a new light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of 40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. Which of the following is the most likely explanation for the statistically significant result?

- a. The new design had more variability than the previous design.
- b. The sample size for the new design is very large.
- c. The mean of 1,200 for the previous design is large.

35. Research participants were randomly assigned to take Vitamin E or a placebo pill. After taking the pills for eight years, it was reported how many developed cancer. Which of the following responses gives the best explanation as to the purpose of randomization in this study?

- a. To reduce the amount of sampling error that can happen if the subjects are not randomly assigned.
- b. To ensure that all potential cancer patients had an equal chance of being selected for the study.
- c. **To produce treatment groups with similar characteristics**
- d. To prevent skewness in the results.

===== **The End** =====



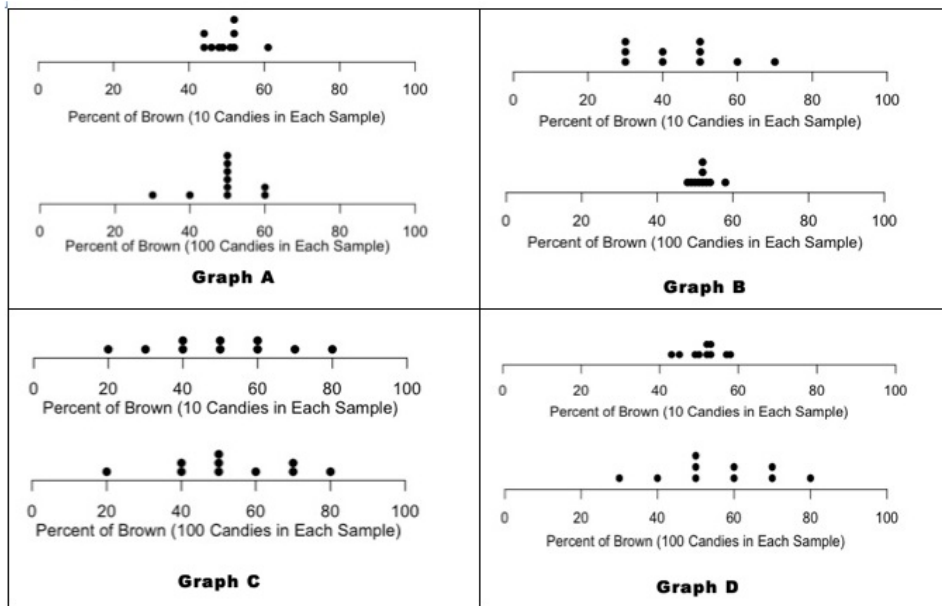
## AIRS-2 (Changes were made from 1st cognitive interview)

### Assessment of Inferential Reasoning in Statistics-2 (AIRS-2)

1. The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched their records for those days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days. The forecast of 70% chance of rain can be considered very accurate if it rained on:

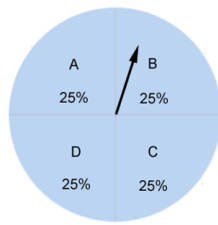
- a. 95% - 100% of those days.
- b. 85% - 94% of those days.
- c. 75% - 84% of those days.
- d. 65% - 74% of those days.
- e. 55% - 64% of those days.

2. Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 50% brown candies. Ten students each take one random sample of 10 candies and record the percentage of brown candies in each of their samples. Another ten students each take one random sample of 100 candies and record the percentage of brown candies in each of their samples. Which of the following pairs of graphs represents the more plausible distributions for the percentage of brown candies obtained in the samples for each group of 10 students?



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

Questions 3 to 8 refer to the following: Consider a spinner shown below that has the letters from *A* to *D*.



'Person 1' used the spinner 10 times and each time he wrote down the letter that the spinner landed on. When he looked at the results, he saw that the letter *B* showed up 5 times out of the 10 spins. Now he doubts the fairness of the spinner because it seems like he got too many *B*s. However, 'Person 2' says that 5 *B*s would not be unusual for this spinner.

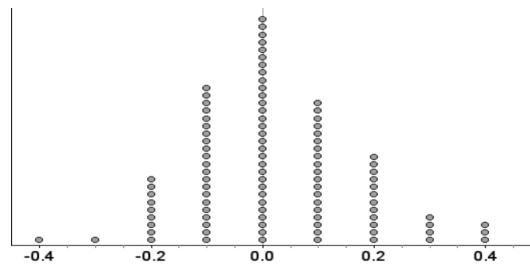
3. If the spinner is fair, how many *B*s out of 10 spins would you expect to see?

- a. 2 or 3 *B*'s
- b. 4 or 5 *B*'s
- c. 6 or 7 *B*'s
- d. 8 or 9 *B*'s

4. A statistician wants to set up a probability model to examine how often the result of 5 *B*'s out of 10 spins could happen with a fair spinner just by chance alone. Which of the following is the best probability model for the statistician to use?

- a. The probability for each letter is the same— $1/4$  for each letter.
- b. The probability for letter *B* is  $1/2$  and the other three letters each have probability of  $1/6$ .
- c. The probability for letter *B* is  $1/2$  and the probabilities for the other letters sum to  $1/2$ .

5. The statistician conducted a statistical test to examine the fairness of the spinner using a computer simulation. The computer simulation randomly generates four letters, A to D. She obtained 100 samples where each sample consisted of 10 letters. She then counted the number of *B*s in each sample of 10 random letters. The following dot plot represents the number of *B*s for each of the 100 samples. What do you think about the observed result of 5 *B*s out of 10 spins in the spinner?



- a. 5 *B*s are not unusual because 5 or less *B*s happened in more than 90 samples out of 100.
- b. 5 *B*s are not unusual because 5 or more *B*s happened in four samples out of 100.
- c. 5 *B*s are unusual because 5 *B*s happened in only three samples out of 100.
- d. 5 *B*s are unusual because 5 or more *B*s happened in only four samples out of 100.
- e. There is not enough information to decide if 5 *B*s are unusual or not.

6. Based on your answers to questions 5 and 6, what would you conclude about whether or not the spinner is fair? Why?

- a. This spinner is most likely fair because 2 Bs and 3 Bs happened the most in the simulation.
- b. This spinner is most likely fair because 5 or less Bs was not unusual in the simulation.
- c. This spinner is most likely unfair because 5 or more Bs was rare in the simulation.
- d. This spinner is most likely unfair because the simulation distribution seems skewed.
- e. We do not know whether or not the spinner is fair because the sample size of 10 is small.

7. Let's say the statistician did another computer simulation, but this time each sample consisted of 20 spins. She calculated the proportion of Bs in each sample (the number of Bs divided by 20). How would you expect the distribution of the *proportion* of Bs obtained from 100 samples of 20 spins each to compare to the distribution of the *proportion* of Bs obtained from 100 samples of 10 spins each?

- a. The distribution of the proportion of Bs for 100 samples of 20 spins each would be wider because you have twice as many spins in each trial.
- b. The distribution of the proportion of Bs for 100 repetitions of 20 spins each would be narrower because you have more information for each sample.
- c. Both distributions would have about the same width because the probability of getting each letter is the same whether you do 10 spins or 20 spins.

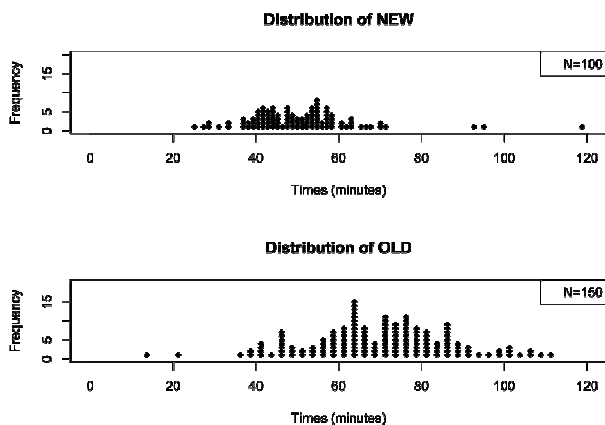
8. Which of the following results, 5 Bs out of 10 spins or 10 Bs out of 20 spins, provides the stronger evidence that the spinner is not fair? Why?

- a. 10 Bs out of 20 spins, because larger samples have less variability, so it is less likely to get an unusual result with a fair spinner.
- b. 5 Bs out of 10 spins, because smaller samples have larger variability, so it is more likely to get an unusual result with a fair spinner.
- c. Both outcomes provide the same evidence because there is the same proportion of Bs ( $1/2$ ) in each of the two samples.

Item 9 to 11 refers to the following situation:

A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown below. Items 9, 10, and 11 present statements made by three different statistics students. For each statement,

indicate whether you think the student's conclusion is valid.



9. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.

- a. Valid
- b. Not valid

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

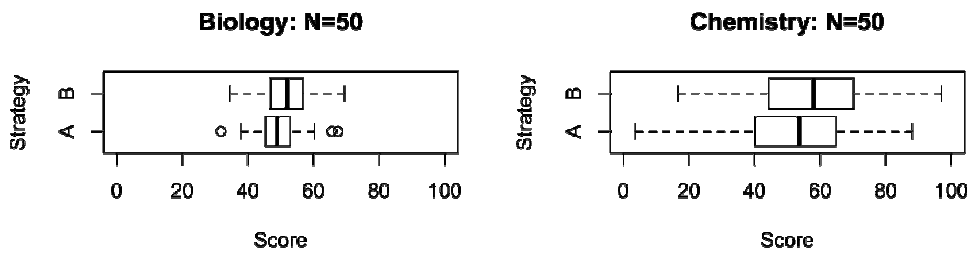
- a. Valid
- b. Not valid

11. We can't conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.

- a. Valid
- b. Not valid

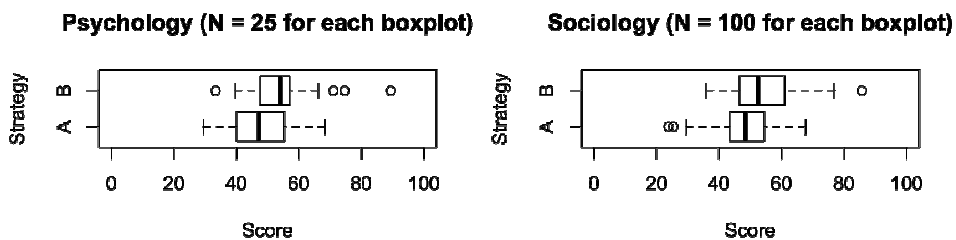
Question 12 and 13 refer to the following: Four experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100) in all four experiments. The four different experiments were conducted with students who were enrolled in four different subject areas: biology, chemistry, psychology, sociology.

12. Boxplots of exam scores for students in the biology course are shown below on the left, and the boxplots for the students in the chemistry course are on the right. For each subject area, 25 students were randomly assigned to either strategy A and 25 students were randomly assigned to strategy B. Which experiment, the one for the biology or the chemistry course, provides the stronger evidence against the claim, "neither strategy is better than the other"?



- Biology, because scores from the Biology experiment are more consistent, which makes the difference between the strategies larger relative to the Chemistry experiment.
- Biology, because the outliers in the boxplot for strategy A from the Biology experiment indicate there is more variability in score for strategy A than for strategy B.
- Chemistry, because scores from the Chemistry experiment are more variable indicating there are more students who got scores above the mean in strategy B.
- Chemistry, because the difference between the maximum and the minimum scores is larger in the Chemistry experiment than in the Biology experiment.

13. Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence against the claim, “neither strategy is better than the other”? Why?

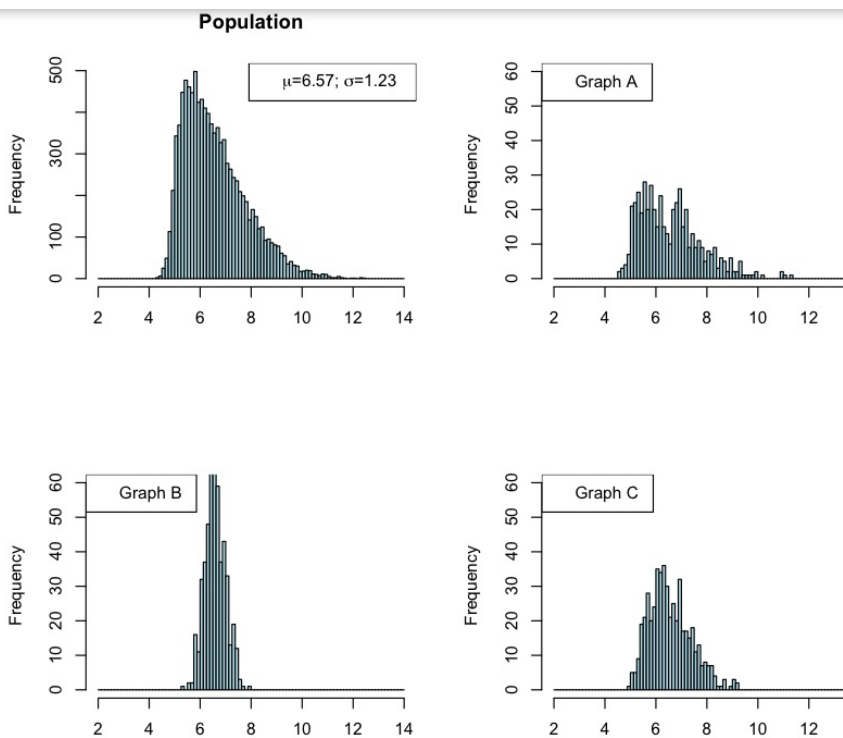


- Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment.
- Psychology, because there are more outliers in strategy B from the Psychology experiment, indicating that strategy B did not work well in that course.
- Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment.
- Sociology, because the sample size is larger in the Sociology experiment, which will produce a more accurate estimate of the difference between the two strategies.

14. A random sample of 25 textbooks for different courses taught at a University is obtained, and the mean textbook price is computed for the sample. To determine the probability of finding another random sample of 25 textbooks with a mean more extreme than the one obtained from this random sample, you would need to refer to:
- the distribution of textbook prices for all courses at the University.
  - the distribution of textbook prices for this sample of University textbooks.
  - the distribution of mean textbook prices for all samples of size 25 from the University.

Questions 15 and 16 refer to the following situation:

Four graphs are presented below. The graph at the top is a distribution for a population of test scores. The mean score is 6.4 and the standard deviation is 4.1.



15. Which graph (A, B, or C) do you think represents a single random sample of 500 values from this population?
- Graph A
  - Graph B
  - Graph C
16. Which graph (A, B, or C) do you think represents a distribution of 500 sample means from random samples each of size 9?
- Graph A
  - Graph B

c. Graph C

17. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?

- a. A random sample of a sample size of 100 with a sample mean of 12.1.
- b. A random sample of a sample size of 36 with a sample mean of 11.5.
- c. A random sample of a sample size of 100 with a sample mean of 11.5
- d. A random sample of a sample size of 36 with a sample mean of 12.1

18. A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes ( $p = 0.287$ ). What does this mean?

- a. The distribution of the GPAs for male and female scholarship athletes are identical except for 28.7% of the athletes.
- b. The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287.
- c. There is a 28.7% chance that a pair of randomly chosen male and female scholarship athletes would have a significant difference assuming that there is no difference.
- d. There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between male and female scholarship athletes as that observed in the sample assuming that there is no difference.

Questions 19 and 20 refer to the following: A researcher investigates the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either be exposed or not be exposed to the herbicide. The fish exposed to the herbicide showed higher levels of an enzyme associated with cancer.

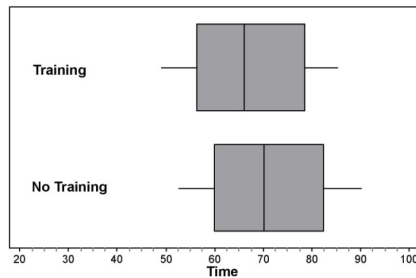
19. Suppose no statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?
- a. The researcher must not be interpreting the results correctly; there should be a significant difference.
  - b. The sample size may be too small to detect a statistically significant difference.
  - c. It must be true that the herbicide does not cause higher levels of the enzyme.

20. Suppose a statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?
- There is evidence of association, but no causal effect of herbicide on enzyme levels.
  - The sample size is too small to draw a valid conclusion.
  - He has proven that the herbicide causes higher levels of the enzyme.
  - There is evidence that the herbicide causes higher levels of the enzyme for these fish.

21 – 22. Read the following information to answer questions 21 and 22:

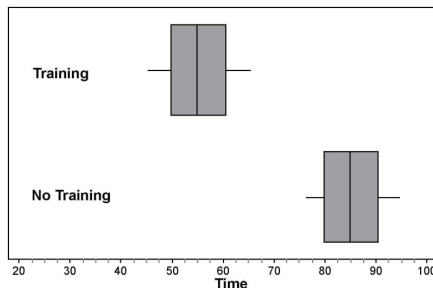
Data are collected from a research study that compares the times to complete a task for professionals who have participated in a new training program with performance for professionals who haven't participated in the program. The professionals are randomly assigned to one of the two groups, with one group receiving the new training program (N=50) and the other group not receiving the training (N=50). For each of the following pairs of graphs, select an appropriate action that you would need to do next to determine if there is a statistically significant difference between the training and no training groups. Write an explanation for your choice.

21.



- Nothing, the two groups appear to be statistically significantly different.
- Conduct an appropriate statistical test for a difference between groups.

22.



- Nothing, the two groups appear to be statistically significantly different
- Conduct an appropriate statistical test for a difference between groups.



23. A student participates in a Coke versus Pepsi taste test. She identifies the correct soda seven times out of ten tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You want to estimate the probability that this student could get *at least seven right out of ten tries just by chance alone*.

You decide to follow a procedure where you:

- Simulate a chance process in which you specify the probability of making a correct guess on each trial
- Repeatedly generate ten cases per trial from this process and record the number of correct outcomes in each trial
- Calculate the proportion of trials where the number of correct guesses meets a specified criterion

In order to run the procedure, you need to decide on the value for the probability of making a correct guess, and specify the criterion for the number of correct guesses.

Which of the options below would provide a reasonable approach to simulating data in order to determine the probability of anyone getting seven out of ten tries correct just by chance alone?

- a. Specify the probability of a correct guess as 50% and calculate the proportion of all trials with *exactly seven* correct guesses
- b. Specify the probability of a correct guess as 50% and calculate the proportion of all trials with *seven or more* correct guesses
- c. Specify the probability of a correct guess as 70% and calculate the proportion of all trials with *exactly seven* correct guesses
- d. Specify the probability of a correct guess as 70% and calculate the proportion of all trials with *seven or more* correct guesses

**Read the following information before answering Questions 24– 26:**

A research question of interest is whether financial incentives can improve performance. Alicia designed a study to test whether video game players are more likely to win on a certain video game when offered a \$5 incentive compared to when simply told to “do your best.” Forty subjects are randomly assigned to one of two groups, with one group being offered \$5 for a win and the other group simply being told to “do your best.” She collected the following data from her study:

	\$5 incentive	“Do your best”	Total
Win	16	8	24
Lose	4	12	16
Total	20	20	40

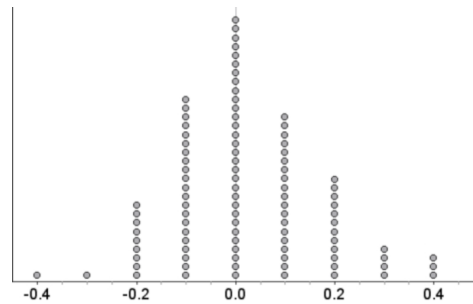
It looks like the \$5 incentive is more successful than the encouragement. The difference in success rates as a proportion is:  $16/20 - 8/20 = 8/20 = 0.40$

In order to test whether this apparent difference might be due simply to chance, she does the following:

- She gets 40 index cards. On 24 of the cards she writes "win" and on 16 she writes "lose".
  - She then shuffles the cards and randomly places the cards into two stacks. One stack represents "\$5 incentive" and the other "verbal encouragement".

- For this simulation, she computes the observed difference in the success rates by subtracting the success rate for the simulation's "\$5 incentive" group from the success rate of the simulation's "verbal encouragement" group.
- She repeats the previous two steps 100 times.
- She plots the 100 statistics she observes from these trials.

The following shows a distribution of simulated data that Alicia generated from her 100 trials and used to test her research question:



24. What is the null model (null hypothesis) that Alicia's data simulated?
- a. The \$5 incentive is more effective than verbal encouragement for improving performance.
  - b. The \$5 incentive and verbal encouragement are equally effective for improving performance.
  - c. Verbal encouragement is more effective than a \$5 incentive for improving performance.
25. Use this distribution to estimate the  $p$ -value for her observed result.
- a. 0.02
  - b. 0.03
  - c. 0.04
  - d. 0.05
  - e. 0.40
26. What does the distribution tell you about the hypothesis that \$5 incentives are effective for improving performance?
- a. The incentive is not effective because the null distribution is centered at 0.
  - b. The incentive is effective because the null distribution is centered at 0.

- c. The incentive is not effective because the  $p$ -value is greater than .05
- d. The incentive is effective because the  $p$ -value is less than .05

Questions 27 to 30 refer to the following: Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached.

27. The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean?

- a. The sample sizes were too small to detect a true difference between the coached and not-coached students.
- b. The observed difference between coached and not-coached students could occur just by chance alone even if coaching really has no effect.
- c. The increase in test scores makes no difference in getting into college since it is not statistically significant.
- d. The study was badly designed because they did not have equal numbers of coached and not-coached students.

28. The study doesn't show that coaching causes a greater increase in college admissions test scores. Which of the following would be the most plausible reason for this?

- a. The not-coached students used other effective ways to prepare.
- b. The number of 4,200 students is too few to detect a difference.
- c. More students were not coached than were coached.

29. The report of the study states, "With 95% confidence, we can say that the average score for students who take the college admissions test a second time is between 28 and 57 points higher than the average score for the first time." By "95% confidence" we mean:

- a. 95% of all students will increase their score by between 28 and 57 points for a second test.
- b. 95% of all students in a new sample will increase their score by between 28 to 57 points for a second test.
- c. 95% of all students who take the college admissions test would believe the statement.
- d. We are 95% certain that the average increase in college admissions scores is between 28 and 57 points.

30. We are 95% confident that the difference between average scores for the first and the second tests is between 28 and 57 points. If we want to be 99% confident, the range of values in the interval would be:

- a. Wider, because higher confidence requires a larger margin of error.
- b. Narrower, because higher confidence requires a smaller margin of error.
- c. Exactly the same width as the range for the 95% confidence interval.

31. A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?

- a. This is a simple random sample. It will give an accurate estimate.
- b. Because the sample is so small, it will not give an accurate estimate.
- c. Because all fans had a chance to be asked, it will give an accurate estimate.
- d. The sampling method is biased. It will not give an accurate estimate.

32. A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that 'prescription medications only' was the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?

- a. Yes, because medication patients lived longer.
- b. No, because doctors chose the treatments.
- c. Yes, because the study was a comparative experiment.
- d. No, because the patients volunteered to be studied.

33. An engineer designs a new light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of 40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. Which of the following is the most likely explanation for the statistically significant result?

- a. The new design had more variability than the previous design.
- b. The sample size for the new design is very large.
- c. The mean of 1,200 for the previous design is large.

34. Research participants were randomly assigned to take Vitamin E or a placebo pill. After taking the pills for eight years, it was reported how many developed cancer. Which of the following responses gives the best explanation as to the purpose of randomization in this study?

- a. To reduce the amount of sampling error that can happen if the subjects are not randomly assigned.
- b. To ensure that all potential cancer patients had an equal chance of being selected for the study.
- c. To produce treatment groups with similar characteristics
- d. To prevent skewness in the results.

===== The End =====

### AIRS-3: Final version (Changes were made from pilot testing)

\*Note: This final version was administered via online assessment tool. This version shown below was copied from the online tool.

#### Assessment of Inferential Reasoning in Statistics (AIRS - 3)

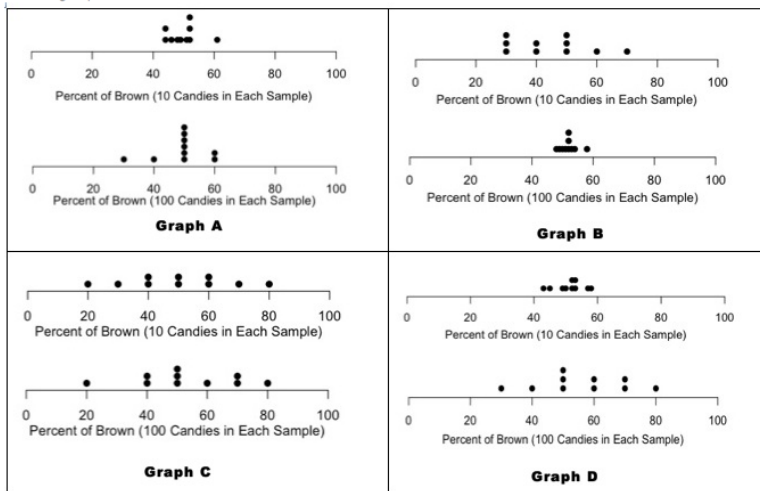
AIRS Online Consent Form

Start AIRS

1. The Springfield Meteorological Center wanted to determine the accuracy of their weather forecasts. They searched their records for those 300 days when the forecaster had reported a 70% chance of rain. They compared these forecasts to records of whether or not it actually rained on those particular days. The forecast of 70% chance of rain can be considered very accurate if it rained on:

- 95% - 100% of those days.
- 85% - 94% of those days.
- 75% - 84% of those days.
- 65% - 74% of those days.
- 55% - 64% of those days.

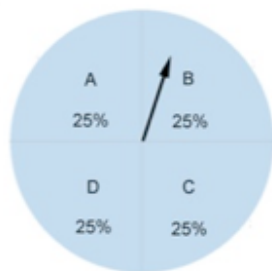
2. Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 50% brown candies. Ten students each take one random sample of 10 candies and record the percentage of brown candies in each of their samples. Another ten students each take one random sample of 100 candies and record the percentage of brown candies in each of their samples. Which of the following pairs of graphs represents the more plausible distributions for the percentage of brown candies obtained in the samples for each group of 10 students?



- Graph A
- Graph B
- Graph C
- Graph D

Questions 3 to 8 refer to the following:

Consider a spinner shown below that has the letters from *A* to *D*.



'Person 1' used the spinner 10 times and each time he wrote down the letter that the spinner landed on. When he looked at the results, he saw that the letter *B* showed up 5 times out of the 10 spins. Now he doubts the fairness of the spinner because it seems like he got too many *B*s. However, 'Person 2' says that 5 *B*s would not be unusual for this spinner.

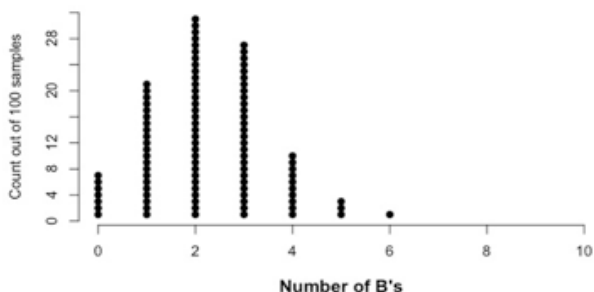
**3. If the spinner is fair, how many *B*s out of 10 spins would you expect to see?**

- 2 or 3 *B*'s
- 4 or 5 *B*'s
- 6 or 7 *B*'s
- 8 or 9 *B*'s

**4. A statistician wants to set up a probability model to examine how often the result of 5 *B*s out of 10 spins could happen with a fair spinner just by chance alone. Which of the following is the best probability model for the statistician to use?**

- The probability for each letter is the same— $1/4$  for each letter.
- The probability for letter *B* is  $1/2$  and the other three letters each have probability of  $1/6$ .
- The probability for letter *B* is  $1/2$  and the probabilities for the other letters sum to  $1/2$ .

**5. The statistician conducted a statistical test to examine the fairness of the spinner using a computer simulation. The computer simulation randomly generates four letters, *A* to *D*. She obtained 100 samples where each sample consisted of 10 letters. She then counted the number of *B*s in each sample of 10 random letters. The following dot plot represents the number of *B*s for each of the 100 samples. What do you think about the observed result of 5 *B*s out of 10 spins in the spinner?**



- 5 *B*s are not unusual because 5 or less *B*s happened in more than 90 samples out of 100.
- 5 *B*s are not unusual because 5 or more *B*s happened in four samples out of 100.
- 5 *B*s are unusual because 5*B*s happened in only three samples out of 100.

- 5 Bs are unusual because 5 or more Bs happened in only four samples out of 100.
- There is not enough information to decide if 5 Bs are unusual or not.

**6. Based on your answers to questions 4 and 5, what would you conclude about whether or not the spinner is fair? Why?**

- This spinner is most likely fair because 2 Bs and 3 Bs happened the most in the simulation.
- This spinner is most likely fair because 5 or less Bs was not unusual in the simulation.
- This spinner is most likely unfair because 5 or more Bs was rare in the simulation.
- This spinner is most likely unfair because the simulation distribution seems skewed.
- We do not know whether or not the spinner is fair because the sample size of 10 is small.

**7. Let's say the statistician did another computer simulation, but this time each sample consisted of 20 spins. She calculated the proportion of Bs in each sample (the number of Bs divided by 20). How would you expect the distribution of the proportion of Bs obtained from 100 samples of 20 spins each to compare to the distribution of the proportion of Bs obtained from 100 samples of 10 spins each?**

- The distribution of the proportion of Bs for 100 samples of 20 spins each would be wider because you have twice as many spins in each trial.
- The distribution of the proportion of Bs for 100 repetitions of 20 spins each would be narrower because you have more information for each sample.
- Both distributions would have about the same width because the probability of getting each letter is the same whether you do 10 spins or 20 spins.

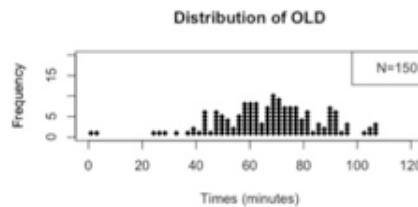
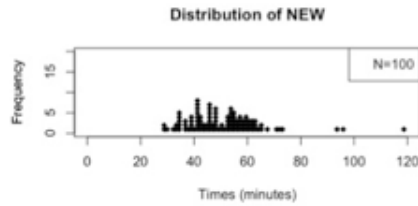
**8. Which of the following results, 5 Bs out of 10 spins or 10 Bs out of 20 spins, provides the stronger evidence that the spinner is not fair? Why?**

- 10 Bs out of 20 spins, because larger samples have less variability, so it is less likely to get an unusual result with a fair spinner.
- 5 Bs out of 10 spins, because smaller samples have larger variability, so it is more likely to get an unusual result with a fair spinner.
- Both outcomes provide the same evidence because there is the same proportion of Bs ( $1/2$ ) in each of the two samples.

---

Item 9 to 11 refers to the following situation:

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



Questions 9, 10, and 11 present statements made by three different statistics students. For each statement, indicate whether you think the student's conclusion is valid.

**9. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.**

- Valid  
 Not valid

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

- Valid  
 Not valid

**11. We can't conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.**

- Valid  
 Not valid

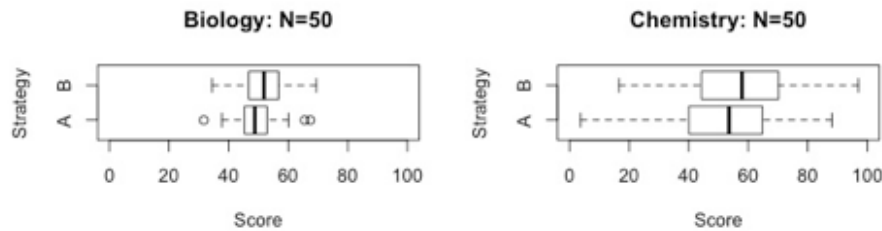
---

Question 12 and 13 refer to the following:

Four experiments were conducted to study the effects of two different exam preparation strategies on exam scores. In each experiment, half of the subjects were randomly assigned to strategy A and half to strategy B. After completing the exam preparation, all subjects took the same exam (which is scored from 0 to 100) in all four experiments. The four different experiments were conducted with students who were enrolled in four different subject areas: biology, chemistry, psychology, sociology.

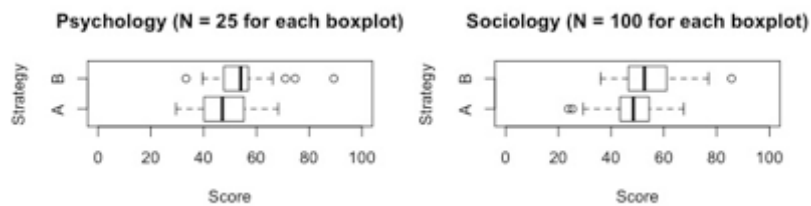


12. Boxplots of exam scores for students in the biology course are shown below on the left, and the boxplots for the students in the chemistry course are on the right. For each subject area, 25 students were randomly assigned to either strategy A and 25 students were randomly assigned to strategy B. Which experiment, the one for the biology or the chemistry course, provides the stronger evidence against the claim, “neither strategy is better than the other”?



- ( ) Biology, because scores from the Biology experiment are more consistent, which makes the difference between the strategies larger relative to the Chemistry experiment.
- ( ) Biology, because the outliers in the boxplot for strategy A from the Biology experiment indicate there is more variability in score for strategy A than for strategy B.
- ( ) Chemistry, because scores from the Chemistry experiment are more variable indicating there are more students who got scores above the mean in strategy B.
- ( ) Chemistry, because the difference between the maximum and the minimum scores is larger in the Chemistry experiment than in the Biology experiment.

**13. Boxplots of exam scores for students in the psychology course are shown below on the left, and the boxplots for the students in the sociology course are on the right. For the psychology course, 25 students were randomly assigned to strategy A and 25 students were randomly assigned to strategy B. However, for the sociology course 100 students were randomly assigned to strategy A and 100 students were randomly assigned to strategy B. Which experiment provides the stronger evidence against the claim, "neither strategy is better than the other"? Why?**



- ( ) Psychology, because there appears to be a larger difference between the medians in the Psychology experiment than in the Sociology experiment.
- ( ) Psychology, because there are more outliers in strategy B from the Psychology experiment, indicating that strategy B did not work well in that course.
- ( ) Sociology, because the difference between the maximum and minimum scores is larger in the Sociology experiment than in the Psychology experiment.

( ) Sociology, because the sample size is larger in the Sociology experiment, which will produce a more accurate estimate of the difference between the two strategies.

---

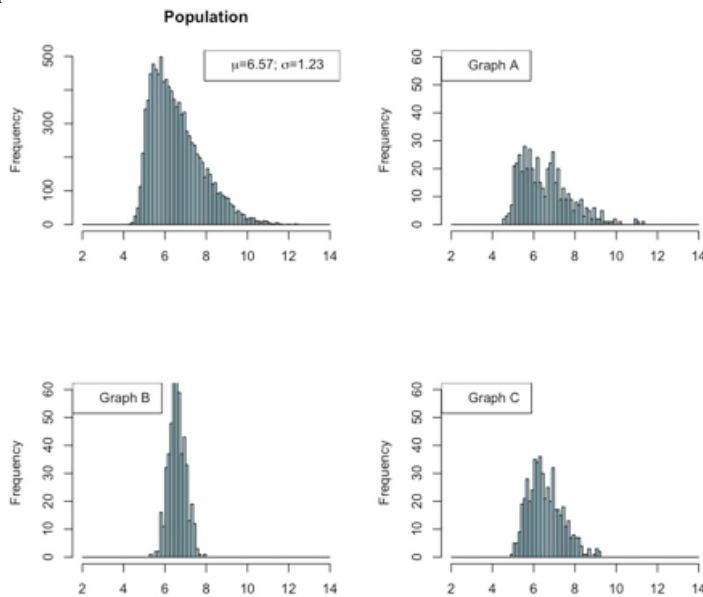
**14. A random sample of 10 textbooks for different courses taught at a University is obtained, and the mean textbook price is computed for the sample. To determine the probability of finding another random sample of 10 textbooks with a mean more extreme than the one obtained from this random sample, you would need to refer to:**

- ( ) the distribution of textbook prices for all courses at the University.
- ( ) the distribution of textbook prices for this sample of University textbooks.
- ( ) the distribution of mean textbook prices for all samples of size 10 from the University.

---

Questions 15 and 16 refer to the following situation:

Four graphs are presented below. The first is a distribution for a population of test scores. The mean score is 6.57 and the standard deviation is 1.23. Please select an appropriate graph for each of the following two questions.



**15. Which graph (A, B, or C) do you think represents a single random sample of 500 values from this population?**

- ( ) Graph A
- ( ) Graph B
- ( ) Graph C

**16. Which graph (A, B, or C) do you think represents a distribution of 500 sample means from random samples each of size 9?**

- Graph A
- Graph B
- Graph C

---

**17. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of adult largemouth bass from Silver Lake. Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?**

- A random sample of sample size of 100 with a sample mean of 12.1.
- A random sample of sample size of 36 with a sample mean of 11.5.
- A random sample of sample size of 100 with a sample mean of 11.5.
- A random sample of sample size of 36 with a sample mean of 12.1.

---

**18. A university administrator obtains a sample of the academic records of past and present scholarship athletes at the university. The administrator reports that no significant difference was found in the mean GPA (grade point average) for male and female scholarship athletes ( $P = 0.287$ ). What does this mean?**

- The distribution of the GPAs for male and female scholarship athletes are identical except for 28.7% of the athletes.
- The difference between the mean GPA of male scholarship athletes and the mean GPA of female scholarship athletes is 0.287.
- There is a 28.7% chance that a randomly chosen male and a randomly chosen female scholarship athlete will have significantly different GPAs assuming that there is no difference.
- There is a 28.7% chance of obtaining as large or larger of a mean difference in GPAs between male and female scholarship athletes as that observed in the sample assuming that there is no difference.

---

Questions 19 and 20 refer to the following:

A researcher investigates the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either be exposed or not be exposed to the herbicide. The fish exposed to the herbicide showed higher levels of an enzyme associated with cancer.

**19. Suppose no statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?**

- ( ) The researcher must not be interpreting the results correctly; there should be a significant difference.
- ( ) The sample size may be too small to detect a statistically significant difference.
- ( ) It must be true that the herbicide does not cause higher levels of the enzyme.

**20. Suppose a statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?**

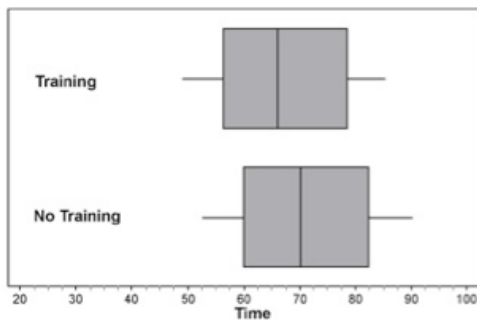
- ( ) There is evidence of association, but no causal effect of herbicide on enzyme levels.
- ( ) The sample size is too small to draw a valid conclusion.
- ( ) He has proven that the herbicide causes higher levels of the enzyme.
- ( ) There is evidence that the herbicide causes higher levels of the enzyme for these fish.

Questions 21 and 22 refer to the following:

Data are collected from a research study that compares the times to complete a task for professionals who have participated in a new training program with performance for professionals who haven't participated in the program. The professionals are randomly assigned to one of the two groups, with one group receiving the new training program (N=50) and the other group not receiving the training (N=50).

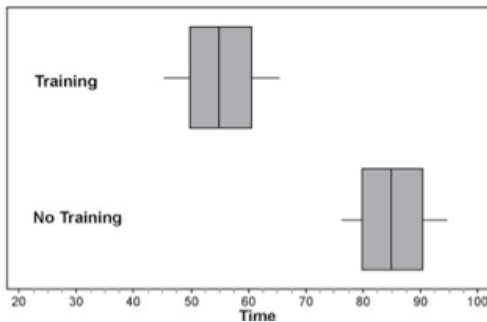
For each of the following pairs of graphs, select an appropriate action that you would need to do next to determine if there is a statistically significant difference between the training and no training groups.

**21.**



- ( ) Nothing, the two groups appear to be statistically significantly different.
- ( ) Conduct an appropriate statistical test for a difference between groups.

**22.**



- ( ) Nothing, the two groups appear to be statistically significantly different.
- ( ) Conduct an appropriate statistical test for a difference between groups.

---

23. A student participates in a Coke versus Pepsi taste test. She correctly identifies the soda seven times out of ten tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You are not sure that she can make this claim. You want to estimate the probability that a student who cannot reliably tell the difference between the two soft drinks could get at least seven right out of ten tries, just by guessing.

You decide to follow a procedure:

1. Simulate a chance process in which you specify the probability of making a correct guess on each trial.
2. Repeatedly generate ten cases per trial from this process and record the number of correct outcomes in each trial.
3. Calculate the proportion of trials where the number of correct guesses meets a specified criterion. In order to run the procedure, you need to decide on the value for the probability of making a correct guess, and specify the criterion for the number of correct guesses.

Which of the options below would provide a reasonable approach to simulating data in order to determine the probability of anyone getting seven out of ten tries correct just by chance alone?

- ( ) Specify the probability of a correct guess as 50% and calculate the proportion of all trials with exactly seven correct guesses.
- ( ) Specify the probability of a correct guess as 50% and calculate the proportion of all trials with seven or more correct guesses.
- ( ) Specify the probability of a correct guess as 70% and calculate the proportion of all trials with exactly seven correct guesses.
- ( ) Specify the probability of a correct guess as 70% and calculate the proportion of all trials with seven or more correct guesses.

---

Questions 24 to 26 refer to the following situation:

A research question of interest is whether financial incentives can improve performance. Alicia designed a study to test whether video game players are more likely to win on a certain video game when offered a \$5 incentive compared to when simply told to "do your best." Forty subjects are randomly assigned to one of two groups, with one group being offered \$5 for a win and the other group simply being told to "do your best." She collected the following data from her study:

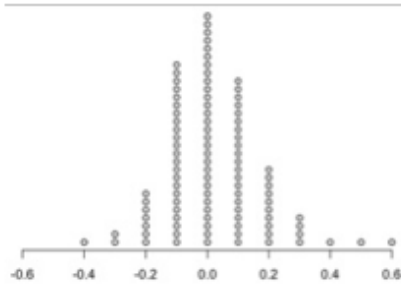
	\$5 incentive	"Do your best"	Total
Win	16	8	24
Lose	4	12	16
Total	20	20	40

It looks like the \$5 incentive is more successful than the encouragement. The difference in success rates as a proportion is:  $16/20 - 8/20 = 8/20 = 0.40$ .

In order to test whether this apparent difference might be due simply to chance, she does the following:

- She gets 40 index cards. On 24 of the cards she writes "win" and on 16 she writes "lose". She then shuffles the cards and randomly places the cards into two stacks. One stack represents "\$5 incentive" and the other "verbal encouragement". For this simulation, she computes the observed difference in the success rates by subtracting the success rate for the simulation's "\$5 incentive" group from the success rate of the simulation's "Do Your Best" (verbal incentive) group.
- She repeats the previous two steps 100 times.
- She plots the 100 statistics she observes from these trials.

The following shows a distribution of simulated data that Alicia generated from her 100 trials and used to test her research question:



24. What is the null model (null hypothesis) that Alicia's data simulated?

- The \$5 incentive is more effective than verbal encouragement for improving performance.
- The \$5 incentive and verbal encouragement are equally effective for improving performance.
- Verbal encouragement is more effective than a \$5 incentive for improving performance.

25. What is the  $P$ -value for her observed result? Use this distribution to estimate the  $P$ -value.

- 0.01
- 0.02
- 0.03
- 0.04
- 0.05

26. What does the distribution tell you about the hypothesis that \$5 incentives are effective for improving performance?

- The incentive is not effective because the null distribution is centered at 0.
- The incentive is effective because the null distribution is centered at 0.
- The incentive is not effective because the  $p$ -value is greater than .05.
- The incentive is effective because the  $p$ -value is less than .05.

Questions 27 to 30 refer to the following:

Does coaching raise college admission test scores? Because many students scored higher on a second try even without coaching, a study looked at a random sample of 4,200 students who took the college admissions test twice. Of these, 500 took a coaching course between their two attempts at the college admissions test. The study compared the average increase in scores for students who were coached to the average increase for students who were not coached.

27. The result of this study showed that while the coached students had a larger increase, the difference between the average increase for coached and not-coached students was not statistically significant. What does this mean?

- The sample sizes were too small to detect a true difference between the coached and not-coached students.
- The observed difference between coached and not-coached students could occur just by chance alone.
- The increase in test scores makes no difference in getting into college since it is not statistically significant.
- The study was badly designed because they did not have equal numbers of coached and not-coached students.

**28. The study doesn't show that coaching causes a greater increase in college admissions test scores. Which of the following would be the most plausible reason for this?**

- The not-coached students used other effective ways to prepare.
- The number of 4,200 students is too few to detect a difference.
- More students were not coached than were coached.

**29. The report of the study states, "With 95% confidence, we can say that the average score for students who take the college admissions test a second time is between 28 and 57 points higher than the average score for the first time." By "95% confidence" we mean:**

- We are certain that 95% of all students will increase their score by between 28 and 57 points for a second test.
- We are certain that 95% of all students in a new sample will increase their score by between 28 to 57 points for a second test.
- We are certain that 95% of all students who take the college admissions test would believe the statement.
- We are 95% certain that the average increase in college admissions scores is between 28 and 57 points.

**30. If we want to be 99% confident that the difference between average scores for the first and the second tests is between 28 and 57 points, the range of values in the interval would be:**

- Wider, because higher confidence requires a larger margin of error.
- Narrower, because higher confidence requires a smaller margin of error.
- Exactly the same width as the range for the 95% confidence interval.

---

**31. A sportswriter wants to know how strongly football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?**

- This is a simple random sample. It will give an accurate estimate.
- Because the sample is so small, it will not give an accurate estimate.
- Because all fans had a chance to be asked, it will give an accurate estimate.
- The sampling method is biased. It will not give an accurate estimate.

---

**32. A study of treatments for angina (pain due to low blood supply to the heart) compared the effectiveness of three different treatments: bypass surgery, angioplasty, and prescription medications only. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. The researchers concluded that 'prescription medications only' was the most effective treatment because those patients had the highest median survival time. Is the researchers' conclusion valid?**

- Yes, because medication patients lived longer.
- No, because doctors chose the treatments.
- Yes, because the study was a comparative experiment.
- No, because the patients volunteered to be studied.

---

**33. An engineer designs a new light bulb. The previous design had an average lifetime of 1,200 hours. The new bulb design has an estimated lifetime of 1,200.2 hours based on a sample of 40,000 bulbs. Although the difference was quite small, the mean difference was statistically significant. A significant result for such a small difference would occur because:**

- The new design had more variability than the previous design.
- The sample size for the new design is very large.
- The mean of 1,200 for the previous design is large.

---

**34. Research participants were randomly assigned to take Vitamin E or a placebo pill. After taking the pills for eight years, it was reported how many developed cancer. Which of the following responses gives the best explanation as to the purpose of randomization in this study?**

- To reduce the amount of sampling error that can happen if the subjects are not randomly assigned.
- To ensure that all potential cancer patients had an equal chance of being selected for the study.
- To produce treatment groups with similar characteristics
- To prevent skewness in the results.

---

Quiz Score

---

Note: Answer key is shown in Appendix K.



## Appendix J

### Expert Review on Preliminary Assessment

Table J-1

*Comments of Reviewers*

Items	Rater	Comments	Rationale for Change Made	Change
4	Internal Reviewer	Remove this item: I could argue for why each response is correct.	All of the responses have all its own argument. All options could be correct.	Item was not removed since we need to see students' actual reasoning.
5	Rater 1	The distracters seem to be very implausible. Might need to have pilot testing using a free-response format.		Changed to free-response question
	Internal Reviewer	"I like this item. However, I would delete the option A. It is not a statement of a probability model. It is a statement about a condition for the trials, which is part of the simulation. Also, in the simulation, you would want the trials to be independent, so it is a correct statement about the simulation"	Agreed	The option A removed.
6	Internal Reviewer	The question is reworded after discussion. The change was made because we decided that students did not quite understand how to simulate the data.	Agreed	Use of 'computer simulation' rather than 'spin more times

*(cont.)*

Items	Rater	Comments	Rationale for Change Made	Change
<i>Table J-1, cont.</i>				
7	Rater 1	Should have another option which says “we don’t know whether the spinner is fair or unfair because...”. In this question, you are setting up two competing hypotheses with the implication that one of them must be accepted but with hypothesis testing all you can do is have evidence against the null (chance alone explanation). If you have no evidence against the null then the two hypotheses remain standing. In other words you do not know whether the spinner is fair or unfair.	Agreed	Added another option, “We do not know whether or not the spinner is fair”.
	Internal Reviewer	Minor wording changes made mostly for the response options made from student interview.		
8	Internal Reviewer	Minor changes to be aligned with item 6.		Use of ‘computer simulation’ rather than ‘spin more times
10	Rater 1	Wording clarification: in option B, include “...on average about 20 minutes sooner than”	Agreed	Included
	Rater 3	I like that the sample sizes are not equal.		
	Internal Reviewer	Item adapted from CAOS. In CAOS, we have these separate items, and the student indicates if they think each statement is Valid or Invalid. You get more information about the students’ thinking if you have them respond to the validity of each statement. You could also then see if a single score based on their responses to all three items provides more information than a separate score for each item	Decided to pilot with three separate items.	Item separated to three.
11	Internal R	Minor wording changes mostly for the response options made from student interview.		

*(cont.)*

Items	Rater	Comments	Rationale for Change Made	Change
<i>Table J-1, cont.</i>				
12	Rater 2	On what informal inference basis are you making a claim? I would pick 'A' using my heuristic.		Decided to leave the original question and see how students are responding in think-aloud.
	Internal R	Minor wording changes mostly for the response options made from student interview.		
13	Rater 3	This is a clunky problem. Do you need to add "of size 25" to part?	Agreed	Added
20	Rater 1	You need to give the sample sizes for both groups and state what the time is measuring. As you state you are comparing two groups since these people are probably volunteers not samples from populations. The learning goal needs to include this idea.	Agreed	Sample size was included. Learning goal was modified.
21	Rater 3	What if n=3 in both groups? Need to add a bit more guidance.	Agreed	Sample size added
23	Rater 3	This is lovely.		
26	Rater 1	You might want to say "observed difference" and "chance alone" for option B.	Agreed	Option B modified
27	Rater 1	Not quite sure if this item is assessing this learning goal. Part of the problem may be that the result was not statistically significant.		
28	Rater 1	Option C should be reworded to better capture ideas about population differences	Agreed	Option c modified
	Rater 3	Wording of option C is clunky and imprecise	Agreed	Option c modified (cont.)

Items	Rater	Comments	Rationale for Change Made	Change
<i>Table J-1, cont.1</i>				
29	Rater 3	Wording comments		Modified
30	Rater 3	Commented about many possible ways to get different answers depending on the proportion of being contaminated of eggs sampled.	Agreed	Item removed
31	Rater 3	Do not think this item gets at the learning goal.	Agreed	Item removed
33	Rater 3	Binomial is less variable when p is close to 0 or 1. Therefore, big differences in true proportions could trump sample size.	Agreed	Item removed
36	Rater 3	I continue to be puzzled why students have such a problem with this item.		

*Note.* Comments of Reviewers: The internal expert's comments were conducted for the revised items from the expert review process and student think-alouds.

## Appendix K

### Reasoning Statement and Expert's Enacted Reasoning

Table K-1

*Reasoning statement (intended reasoning) in AIRS-1*

Item #	Correct Answer	Intended Reasoning
1 Forecast	D	Since it is reported 70 % chance of raining, the interval for the population proportion of raining should include 70%.
2 Brown candies	B	The proportions of the brown candies in ten candies will be more closely clustered to the mean proportion (.5) for 100 samples than for 10 samples because smaller samples tend to have larger variability.
3 Spinner 1: How many B's you expect	A	If the spinner is fair, the number of letters being landed would be equally likely. Since there are four possibilities, each of the letters has the equal chance of a quarter—about two or three spins out of 10.
4 Spinner 2: Null model	A	The null hypothesis is the one that will happen assuming the spinner is fair: each letter has an equal change of a quarter.
5 Spinner 3: distribution of 100 samples	D	5 Bs out of 10 spins is unusual if the spinner is fair, because from the distribution of 100 samples, there are only 4 cases where 5 Bs or more Bs happened out of 10 spins.
6 Spinner 4: Is the spinner fair?	C	This spinner is not fair because from the distribution above we observed that 5 Bs out of 10 spins happened only 4 times when the spinner is fair.
7 Spinner 5: 20 samples	B	The distribution of the proportion of Bs obtained from 100 samples of 20 spins would be narrower because there would be less variability in a larger sample size.
8 Spinner 7: which one is the stronger evidence?	A	Since the 100 samples of 20 spins have narrower distribution than 10 spins, it would be less likely to get an unusual result with a fair spinner. Therefore, 100 samples of 20 spins would be the stronger evidence to support that the spinner is not fair.
9 A drug company 1	B	Invalid. We need to see in which group <i>chunk</i> of people have less time to get relief. This statement focuses only on some of the data, not about the general tendency of the data. (Students are expected to see the data as aggregates not as individual data)
10 A drug company 2	A	Valid because the average time for the new formula group is larger.

*(cont.)*

Item #	Correct Answer	Intended Reasoning
<i>Table K-1, cont.</i>		
11 A drug company 3	B	Invalid. Although the sample sizes are different for two groups, we can make a conclusion because both sample sizes are fairly large.
12 Exam strategy 1	A	The sample size and mean difference between two strategies look the same in Biology and Chemistry. However, Biology has narrower distribution meaning it has smaller variability than Chemistry. This indicates that the difference between two groups is more consistent (or reliable), so it has stronger evidence that there is a difference between two groups.
13 Exam strategy 2	D	The variability and a difference between two strategies look similar in Psychology and Sociology. However, Sociology has a larger sample indicating the sample of Sociology is more representative to the population.
14 Textbook	C	Since we want to know how expensive the sample of 25 textbooks is, we need a sampling distribution of all samples of size 25 from the population (university).
15 A single random sample of 500	A	A single random sample of 500 values would be representative of a population.
16 500 sample means	B	A distribution of 500 sample means would follow the Central Limit Theorem—normally distributed centered to the mean, less variability.
17 Silver Lake fish	C	The smaller sample and the larger the sample size, the stronger evidence.
18 GPA	D	Interpretation of the p-value of 28.7%.
19 Herbicide to fish: no statistical significance	B	It is possible that a statistical testing could not capture the observed difference because of small sample size.
20 Herbicide to fish: a statistical significance	D	Since the fish were randomly assigned to two groups, we can make a causal inference from the statistical significant result.
21 Training vs. No-training with overlaps	B	Since there is an overlap between two groups, we need to do a statistical test to see if the difference indicates a <i>statistically significant difference</i> .

(cont.)

Item #	Correct Answer	Intended Reasoning
<i>Table K-1, cont.</i>		
22 Training vs. No-training without overlaps	A	Since there is no overlap between two groups, we can conclude that there is a significant difference.
23 Coke vs. Pepsi	B	The probability of guessing is 50% and what we observed in our sample is seven out of ten. Therefore, 50% of chance would be the probability of specification and calculate the proportion of all trials with seven or more correct guesses.
24 Alicia, null model	B	The null model is one that we have the result just by chance. Therefore, null model here is that there is equally likely effectiveness.
25 Alicia, p-value	B or C	Since we have found four times out of 100 where the cases are greater than the observed proportion of 0.4, the p-value is 0.03 (or 0.04 if we consider both sides).
26 Alicia, conclusion	D	Since the p-value is less than 0.05, we reject the null. The incentive is effective.
27 coaching – no statistical significance	B	Since the sample size is large enough and there was no significant difference between two groups, the observed difference could happen just by chance alone.
28 coaching – statistical significance	A	Since there was no random assignment for treatment, any confounding factors could've have impact on the observed result.
29 95% CI	D	The confidence interval indicates the range of increase score in a second test for the population. This gives us the degree of certainty.
30 Range of 99% CI	A	If the confidence level increases, the margin of error increases. Therefore, the range of values gets wider.
31 sports writer	D	This is a biased sampling because the sample (people who went to the football stadium) is not representative to a population.
32 angina	B	This is an experiment with no random assignment. The conclusion is not valid because the doctors chose the treatment groups.
33 bulb	B	Since the sample size is very large, even a small observed difference could result in a statistically significant difference.
34 Vitamin vs. placebo	C	The purpose of random assignment is to have equal characteristics for both of treatment group and control group.

Table K-2

*A Script from an Expert's Think-aloud*

Items	Intended Reasoning	Enacted Reasoning (expert's reasoning)
Item 5: Spinner 3: Null model	The null hypothesis is the one that happened if the spinner is fair.	Since we have 10 spins, and we want to have a probability model, and we want to count the number of B's, based on the set-up of the spinner, it looks like each letter has equal probability of being chosen, and because it's fair. The probability model is gonna be based on the fair spinner. Each letter would have to have equal probability. If I would spin the fair spinner ninety times, not just ten. This fair spinner in the long run, the probability of each letter would come out to be about one quarter.
Item 9-11: A drug company	Invalid. We need to see in which group <i>chunk</i> of people have less time to get relief. This statement focuses only on some of the data, not about the general tendency of the data. (Students are expected to see the data as aggregates not as individual data)	This statement is not valid. Because it looks to me like...if you look at the overall shape of this data, the overall average of old formula would be larger than the overall average of the new formula, which means that the new formula works better.
Item 10.	Valid because the average time for the new formula group is larger.	I agree with the first statement. And on average makes sense to me. So I would say it's valid.
Item 11.	Invalid. Although the sample sizes are different for two groups, we can make a conclusion because both sample sizes are fairly large.	That is not valid. Two groups were chosen randomly, the number of samples is fairly large, so I think we can make some conclusion on the comparison.
Item 12-13. Biology and Chemistry: Item 12.	Since the sample size and a difference between two samples look the same, we need to look at the distribution of two. Biology has narrower distribution indicating that the difference between two groups is more consistent (or reliable), so it has stronger evidence that there is a difference between two groups.	In both of the box plots, the boxes overlap quite significantly. And the tails are also overlap. The chemistry, there are same amount of variability between two strategies. And the biology, there are less variations than the chemistry for both strategies. So I would say the less variability means the scores are more consistent in Biology. Given that the difference between two strategies is almost the same in two groups (Biology and Chemistry) the less variability gives stronger evidence against the claim.

*(cont.)*



Items	Intended Reasoning	Enacted Reasoning (expert's reasoning)
<i>Table K-2, cont.</i>		
Item 18.	Interpretation of the p-value of 28.7%.	It's basically asking about the definition of p-value. So I would say D is the correct answer.
Item 19.	If there is no statistical difference between two groups of fish in an experiment where they found some difference, it could be because of a small sample size.	I don't think it's A because they say that it is statistically significant. I would say B is correct: the same size is sixty. If we have more fish, he could have better idea of what the difference of two groups, it might tell better.
Item 20.	If there is a statistical difference between two groups of fish in an experiment with random assignment, it indicates that we have evidence of causation.	I did random assignment. So, not A. Possible for B, but he found significant difference, so not B. I would say D instead of C. because the idea of having evidence causes higher levels of the enzyme given that we used the random assignment. Even so, we couldn't say we could prove something.
Item 24-26.	The null model is one that we have the result just by chance. Therefore, no improvement with \$5 incentive.	Her null model is based on the fact that they are equally effective. So, I would say the answer is B showing both of the groups are equally effective for the performance.
Item 25	Since we have found four times out of 100 which is great than 0.4, the p-value is 0.03 (or 0.04 if we consider both sides)	She's taking the difference between. I see that she only cares one-sided where or not there is improvement. So, it's three out of 100.
Item 26	Since the p-value is less than 0.05, we reject the null. The incentive is effective.	Since the p-value is less than 0.05, so I would say the incentive is effective.
Item 27.	Since the sample size is large enough and there was no significant difference between two groups' scores, the observed difference could happen just by chance alone.	I would say sample size is fairly large, so A is not the answer. I would say B, because we did see a difference but it wasn't significant. That means that happened just by chance alone even if coaching really has not any effect.
Item 28.	This is an experiment study with no random assignment. If there was not a significant difference between two groups, it could be because any confounding factors were not controlled.	I would say that there are any effective ways to prepare for the not-coached students. That makes the most sense to me.

(cont.)

Items	Intended Reasoning	Enacted Reasoning (expert's reasoning)
<i>Table K-2, cont.</i>		
Item 29.	The confidence interval indicates the range of increase score in a second test for the population. This gives us the degree of certainty.	95% CI means just D. this is about the definition of confidence interval.
Item 33.	Since the sample size is very large, the small observed difference could be compensated to be statistically significant.	I would say the answer is B, because with huge sample size like this we can get a significant result even with a tiny difference between two groups.
Item 34.	The purpose of random assignment is to control any confounding factors by having all subjects be selected with an equal chance.	This is basically asking about the purpose of random assignment. If you are randomly assigning the people to two groups, Vitamin and placebo, we can even out the systematic difference between them. So B is the most plausible answer because this way (random assigning) any difference within or between groups can be controlled.

*Note.* The think-aloud with an expert was conducted before the 1<sup>st</sup> cognitive interview.

## Appendix L

### Reliability Analysis from Pilot Testing

Item	Standardized Alpha	Polyserial Correlation
1	0.82	0.86
2	0.83	0.84
3 <sup>a</sup>	NA	NA
4	0.84	-0.27
5	0.82	0.9
6	0.84	0.53
7	0.83	0.61
8	0.83	0.63
9 <sup>a</sup>	NA	NA
10	0.83	0.54
11	0.83	0.71
12	0.83	0.37
13	0.84	0.12
14	0.83	-0.12
15	0.82	0.66
16	0.82	0.59
17	0.83	0.59
18	0.83	0.65
19	0.84	0.18
20	0.84	0.03
21	0.83	0.51
22	0.84	0.12
23	0.84	0.27
24	0.83	0.31

*(cont.)*

Item	Standardized Alpha	Polyserial Correlation
<i>Table L, cont.</i>		
25	0.84	0.21
26	0.84	0.29
27	0.82	0.77
28	0.83	0.74
29	0.84	-0.14
30	0.83	0.53
31	0.83	0.77
32	0.82	0.64
33	0.82	1
34	0.84	0.56
Total standardized alpha = 0.84		

<sup>a</sup>Item 3 and item 9 have perfect correct score, so coefficient alpha and item-total correlation are not available.

## Appendix M

### LD Indexes of AIRS Items

Note: The lower diagonal presents Likelihood Ratio  $G^2$  statistic for each pair of 34 items. The upper diagonal shows Cramer's V.

294

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
Q1		0.05	0.02	0.03	0.02	0.03	0.05	0.02	0.04	0.01	0.02	0.00	0.04	0.01	0.03	0.02	0.04	0.02	0.05
Q2	5.15		0.04	0.05	0.01	0.04	0.19	0.01	0.07	0.08	0.05	0.01	0.03	0.02	0.03	0.01	0.04	0.01	0.04
Q3	0.87	-2.90		0.08	0.03	0.03	0.06	0.02	0.02	0.03	0.04	0.02	0.03	0.03	0.03	0.03	0.03	0.04	0.03
Q4	1.81	-4.11	13.78		0.03	0.01	0.04	0.04	0.03	0.03	0.02	0.01	0.00	0.03	0.01	0.03	0.04	0.05	0.02
Q5	-0.88	0.10	-1.31	-1.59		0.07	0.02	0.01	0.02	0.01	0.04	0.04	0.01	0.03	0.00	0.04	0.02	0.03	0.01
Q6	-1.26	-2.39	-1.65	-0.23	9.33		0.01	0.04	0.03	0.04	0.02	0.01	0.04	0.02	0.01	0.04	0.02	0.05	0.03
Q7	5.65	69.60	-6.02	-2.80	-0.58	-0.05		0.09	0.07	0.09	0.05	0.03	0.03	0.01	0.03	0.01	0.08	0.02	0.02
Q8	0.92	-0.14	0.98	-2.84	0.20	-3.34	16.76		0.04	0.01	0.01	0.00	0.00	0.05	0.03	0.02	0.04	0.02	0.02
Q9	-3.69	-9.34	0.92	1.43	-0.72	-1.33	-10.05	-2.58		0.11	0.06	0.04	0.05	0.03	0.03	0.02	0.02	0.06	0.02
Q10	0.07	-13.82	-1.42	-1.29	0.38	-3.55	-15.44	-0.28	23.88		0.20	0.03	0.03	0.05	0.01	0.01	0.04	0.04	0.01
Q11	-0.64	-4.99	-2.91	-0.56	-3.86	-0.86	-5.68	-0.24	-6.43	81.22		0.02	0.05	0.02	0.04	0.02	0.05	0.01	0.04
Q12	0.01	-0.07	1.10	0.26	-3.00	0.26	-1.68	-0.01	-2.78	-1.55	-1.13		0.06	0.01	0.01	0.01	0.02	0.02	0.01
Q13	-2.98	-1.34	1.41	0.02	0.05	-2.42	-1.86	0.01	4.21	-2.02	-5.48	6.23		0.01	0.06	0.04	0.07	0.04	0.01
Q14	-0.31	0.57	-2.04	-1.24	-2.00	0.85	0.14	-3.93	-1.20	-4.06	-0.72	0.11	0.23		0.03	0.05	0.03	0.03	0.07
Q15	1.66	2.31	-1.50	-0.19	-0.01	0.15	-1.22	-2.11	-1.94	-0.13	-2.71	0.38	-6.12	1.61		0.17	0.04	0.03	0.01
Q16	0.98	0.07	1.28	-1.60	-2.57	-3.61	-0.27	-1.10	-0.94	-0.20	-1.01	0.05	-3.72	5.65	56.06		0.01	0.02	0.00
Q17	-2.62	-3.41	-2.19	-3.06	-0.42	-0.70	-11.13	-2.50	-0.89	-2.62	-4.10	-0.45	10.19	-2.07	-3.12	0.40		0.02	0.01
Q18	-0.93	-0.07	-3.39	-4.78	2.08	4.31	-0.62	-0.76	-6.61	-3.88	0.39	-0.43	-2.43	1.27	1.61	0.97	0.70		0.01
Q19	-4.15	-3.18	-1.32	1.08	0.05	-1.34	-0.76	-0.91	-0.70	0.11	-3.17	0.18	-0.24	-8.69	-0.36	0.01	-0.39	0.28	

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
Q20	-4.91	-4.62	-0.51	0.04	1.08	0.43	-8.14	-3.69	2.43	0.22	0.31	-1.58	0.16	-0.41	-0.18	-11.14	5.24	-1.02	-4.94
Q21	-2.21	-1.62	0.85	1.10	-4.51	0.34	-16.60	-5.74	-2.12	-3.78	-1.11	-0.41	0.41	-2.81	-0.76	-0.53	-2.35	-2.23	1.52
Q22	0.63	-2.08	12.63	4.49	-1.09	-1.09	-10.79	0.00	0.79	2.01	-0.71	-0.01	0.05	-5.66	-0.03	-0.10	0.44	-3.86	-0.03
Q23	-1.50	2.56	-2.48	2.14	6.97	0.47	-0.02	-0.20	-3.52	0.43	0.06	-0.63	-0.81	-0.19	-1.71	-0.40	-3.11	-0.19	0.60
Q24	3.20	-1.43	-2.74	-2.40	0.17	-0.39	0.63	-0.68	-10.00	-3.46	-1.86	0.13	-1.11	-0.51	0.00	0.92	-1.04	-0.47	-0.43
Q25	0.33	0.48	-0.58	0.04	-1.27	12.29	-0.07	-0.02	-11.27	-9.67	-1.18	-0.04	-0.04	-3.17	1.91	-0.45	-5.62	0.20	7.21
Q26	-4.51	-2.86	0.51	-0.82	0.01	4.89	-5.25	-3.12	1.48	-2.00	-0.02	0.62	-1.08	-0.03	1.02	-0.16	-0.52	1.26	-1.29
Q27	-1.87	-2.21	-3.84	-0.77	-0.26	-0.03	-0.56	-0.24	-8.38	-0.84	5.29	-0.10	-12.28	-0.07	-1.31	-1.65	-5.01	-3.43	-0.04
Q28	-0.18	-3.03	-4.20	-3.95	-0.20	-0.25	-0.02	-0.26	-5.55	0.18	8.50	-1.18	-8.25	-0.46	-5.00	-2.55	-8.03	-0.10	-5.97
Q29	-5.18	-0.56	-2.54	1.42	-4.15	-0.19	-1.86	-4.76	-1.32	-1.69	0.09	-0.54	1.00	-4.86	-0.08	-5.94	0.43	0.05	-0.35
Q30	4.64	2.13	-0.53	-0.55	-0.02	-0.62	2.41	-2.04	-8.58	-3.89	-2.77	-3.92	-4.99	0.29	0.04	-0.06	-7.62	-0.32	0.16
Q31	-1.68	-6.40	-1.25	-1.09	-4.88	-1.02	-3.21	-3.14	-1.54	-4.05	-10.50	-0.43	-0.11	0.97	-0.72	1.51	-1.98	-1.29	0.35
Q32	-0.11	-0.27	-0.77	-3.10	-0.08	-3.20	1.26	-1.56	-2.11	-4.30	-14.78	6.43	-0.04	0.53	3.83	5.97	-3.23	0.41	0.61
Q33	0.24	-0.72	-2.47	0.22	0.56	1.24	0.50	0.90	-6.75	-3.33	-4.90	-0.02	6.44	-2.55	-0.56	1.50	-1.84	-0.51	5.27
Q34	0.92	1.01	-13.83	-0.86	0.07	15.19	0.52	-0.29	-20.85	-4.61	-0.23	4.57	-3.01	0.04	3.43	-0.03	-5.59	5.13	-4.30

	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34
Q1	0.05	0.03	0.02	0.03	0.04	0.01	0.05	0.03	0.01	0.05	0.05	0.03	0.01	0.01	0.02
Q2	0.05	0.03	0.03	0.04	0.03	0.02	0.04	0.03	0.04	0.02	0.03	0.06	0.01	0.02	0.02
Q3	0.02	0.02	0.08	0.04	0.04	0.02	0.02	0.04	0.05	0.04	0.02	0.03	0.02	0.04	0.08
Q4	0.00	0.02	0.05	0.03	0.04	0.01	0.02	0.02	0.05	0.03	0.02	0.02	0.04	0.01	0.02
Q5	0.02	0.05	0.02	0.06	0.01	0.03	0.00	0.01	0.01	0.05	0.00	0.05	0.01	0.02	0.01
Q6	0.02	0.01	0.02	0.02	0.01	0.08	0.05	0.00	0.01	0.01	0.02	0.02	0.04	0.03	0.09
Q7	0.06	0.09	0.07	0.00	0.02	0.01	0.05	0.02	0.00	0.03	0.04	0.04	0.03	0.02	0.02
Q8	0.04	0.05	0.00	0.01	0.02	0.00	0.04	0.01	0.01	0.05	0.03	0.04	0.03	0.02	0.01
Q9	0.04	0.03	0.02	0.04	0.07	0.08	0.03	0.07	0.05	0.03	0.07	0.03	0.03	0.06	0.10
Q10	0.01	0.04	0.03	0.02	0.04	0.07	0.03	0.02	0.01	0.03	0.04	0.05	0.05	0.04	0.05
Q11	0.01	0.02	0.02	0.01	0.03	0.02	0.00	0.05	0.07	0.01	0.04	0.07	0.09	0.05	0.01
Q12	0.03	0.01	0.00	0.02	0.01	0.01	0.02	0.01	0.02	0.02	0.05	0.02	0.06	0.00	0.05
Q13	0.01	0.01	0.01	0.02	0.02	0.00	0.02	0.08	0.07	0.02	0.05	0.01	0.01	0.06	0.04
Q14	0.01	0.04	0.05	0.01	0.02	0.04	0.00	0.01	0.02	0.05	0.01	0.02	0.02	0.04	0.01
Q15	0.01	0.02	0.00	0.03	0.00	0.03	0.02	0.03	0.05	0.01	0.01	0.02	0.04	0.02	0.04
Q16	0.08	0.02	0.01	0.01	0.02	0.02	0.01	0.03	0.04	0.06	0.01	0.03	0.06	0.03	0.00
Q17	0.05	0.03	0.02	0.04	0.02	0.05	0.02	0.05	0.06	0.02	0.06	0.03	0.04	0.03	0.05
Q18	0.02	0.03	0.04	0.01	0.02	0.01	0.03	0.04	0.01	0.01	0.01	0.03	0.01	0.02	0.05
Q19	<b>0.05</b>	0.03	0.00	0.02	0.02	0.06	0.03	0.01	0.06	0.01	0.01	0.01	0.02	0.05	0.05
Q20		0.04	0.04	0.03	0.04	0.06	0.03	0.02	0.01	0.04	0.02	0.03	0.09	0.02	0.03
Q21	2.50		<b>0.40</b>	0.03	0.02	0.02	0.03	0.05	0.02	0.03	0.06	0.02	0.03	0.03	0.05
Q22	3.28	<b>311.87</b>		0.04	0.03	0.02	0.00	0.08	0.01	0.01	0.04	0.01	0.03	0.00	0.07
Q23	-1.37	-2.17	-2.58		0.02	0.03	0.00	0.02	0.01	0.02	0.05	0.09	0.02	0.01	0.02

	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34
Q24	-3.20	0.82	-1.42	0.88		<b>0.01</b>	<b>0.01</b>	0.05	0.01	0.03	0.01	0.08	0.02	0.05	0.02
Q25	-6.21	-1.18	-1.18	1.46	<b>-0.24</b>		<b>0.14</b>	0.00	0.04	0.01	0.02	0.08	0.01	0.01	0.06
Q26	1.38	1.39	0.01	-0.01	<b>-0.29</b>	<b>38.38</b>		0.03	0.01	0.04	0.06	0.01	0.02	0.03	0.01
Q27	0.88	-5.08	-11.45	0.65	4.39	0.03	-1.85		<b>0.22</b>	<b>0.06</b>	<b>0.01</b>	0.03	0.00	0.03	0.00
Q28	0.13	0.45	-0.12	-0.07	0.18	-3.37	-0.20	<b>92.21</b>		<b>0.01</b>	<b>0.04</b>	0.02	0.04	0.07	0.01
Q29	3.17	2.02	0.26	-0.50	-2.11	0.21	2.98	<b>-7.76</b>	<b>0.41</b>		<b>0.00</b>	0.01	0.00	0.03	0.00
Q30	-0.60	-6.03	-3.21	-4.88	-0.24	-0.48	-8.01	<b>0.07</b>	<b>3.32</b>	<b>-0.03</b>		0.03	0.04	0.03	0.04
Q31	1.64	-1.16	-0.11	-16.80	-12.34	-13.82	0.20	-1.47	-0.86	0.18	1.69		0.04	0.02	0.07
Q32	-17.05	-1.59	-1.51	-1.08	0.56	0.40	1.01	0.04	-2.58	-0.03	2.37	2.49		0.01	0.06
Q33	-0.68	-1.36	0.00	0.25	4.91	0.28	-1.56	-1.23	-10.52	-1.56	-1.64	-0.52	-0.16		0.02
Q34	-1.79	-4.11	-9.68	0.44	1.01	7.99	-0.38	-0.02	0.26	-0.01	3.34	-9.64	5.91	-0.49	



## Appendix N

### Development of a Preliminary Version:

#### Item Changes Made from Existing Instruments

Item Numbers in Preliminary Version AIRS	Item Source and Original Item	Changes Made and Rationale for Change
1	Konold and Garfield (1993), as adapted from Falk 1993, problem 5.1.1, p. 111	No change
2	Context adapted from CAOS item 17.	Item was revised by the author to ask: - Understanding the nature and behavior of sampling variability - Understanding sample to sample variability - Taking into account sample size in association with sampling variability
3-9 Spinner problem	CATALST project (ongoing validation) items: [Context omitted] Q. How could you decide which person is correct? Explain. Q. Did you use technology to answer this question? If so please describe what you used. Explain what you think this p value suggests about whether or not the spinner is fair? Q. Do you think this result would produce the same $p$ -value of 0.08 as before, or a higher $p$ -value, or a lower one? Explain your reasoning. Q. Did you use technology to answer questions 3 or 4? If so please describe what you used.	The scenario of the items was adopted and revised. The items were revised to MC types. The items were created by the author and delMas.

*(cont.)*

Item Numbers in Preliminary Version AIRS	Item Source and Original Item	Changes Made and Rationale for Change
10	<p>CAOS item 11- 13: [Context omitted]</p> <p>11. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.</p> <p>a. Valid. b. Not valid.</p> <p>12. The average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.</p> <p>a. Valid. b. Not valid.</p> <p>13. I would not conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.</p> <p>a. Valid. b. Not valid.</p>	<p>The original three items in CAOS was merged to one item.</p>
11, 12	<p>Context adapted from CATALST project (ongoing validation)</p> <p>Items created by Robert delMas on the topic of Comparing two samples from two populations</p>	<i>(cont.)</i>

Item Numbers in Preliminary Version AIRS	Item Source and Original Item	Changes Made and Rationale for Change
<i>Table N, cont.</i>		
13	<p data-bbox="394 578 926 605">ARTIST topic scale (Sampling Variation) item 4:</p> <p data-bbox="394 623 1251 740">A random sample of 25 college statistics textbook prices is obtained and the mean price is computed. To determine the probability of finding a more extreme mean than the one obtained from this random sample, you would need to refer to:</p> <ul style="list-style-type: none"> <li data-bbox="394 758 1121 786">a. the population distribution of all college statistics textbook prices.</li> <li data-bbox="394 787 1171 815">b. the distribution of prices for this sample of college statistics textbooks.</li> <li data-bbox="394 816 1220 878">c. the sampling distribution of textbook prices for all samples of 25 textbooks from this population.</li> </ul>	<p data-bbox="1289 592 1829 805">14. A random sample of 10 textbooks for different courses taught at a University is obtained, and the mean textbook price is computed for the sample. To determine the probability of finding another random sample of 10 textbooks with a mean more extreme than the one obtained from this random sample, you would need to refer to:</p> <ul style="list-style-type: none"> <li data-bbox="1289 816 1745 878">a. the distribution of textbook prices for all courses at the University.</li> <li data-bbox="1289 880 1759 941">b. the distribution of textbook prices for this sample of University textbooks.</li> <li data-bbox="1289 943 1814 1000">c. the distribution of mean textbook prices for all samples of size 10 from the University.</li> </ul>
14, 15	CAOS 34, 35	No change

Item Numbers in Preliminary Version AIRS	Item Source and Original Item	Changes Made and Rationale for Change
16	<p>CAOS 32:            [Context omitted] A research group from the Department of Natural Resources took a random sample of 100 adult largemouth bass from Silver Lake and found the mean of this sample to be 11.2 inches. Which of the following is the most appropriate statistical conclusion?</p> <p>a. The researchers cannot conclude that the fish are smaller than what is normal because 11.2 inches is less than one standard deviation from the established mean (12.3 inches) for this species.</p> <p>b. The researchers can conclude that the fish are smaller than what is normal because the sample mean should be almost identical to the population mean with a large sample of 100 fish.</p> <p>c. The researchers can conclude that the fish are smaller than what is normal because the difference between 12.3 inches and 11.2 inches is much larger than the expected sampling error.</p>	<p>Used the same context but modified in wording and alternatives:            17.[Context omitted] Which of the following provides the strongest evidence to support the claim that they are catching smaller than average length (12.3 inches) largemouth bass this year?</p> <p>a. A random sample of a sample size of 100 with a sample mean of 12.1.</p> <p>b. A random sample of a sample size of 36 with a sample mean of 11.5.</p> <p>c. A random sample of a sample size of 100 with a sample mean of 11.5.</p> <p>d. A random sample of a sample size of 36 with a sample mean of 12.1.</p>
17	Adapted from Instructor's Manual and Test Bank for Moore and Notz' (Moore et al., 2008)	<i>(cont.)</i>

Item Numbers in Preliminary Version AIRS	Item Source and Original Item	Changes Made and Rationale for Change
18, 19	CAOS 23, 24: A researcher in environmental science is conducting a study to investigate the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either a treatment or a control group. The fish in the treatment group showed higher levels of the indicator enzyme.	Change in wording of the context and questions to make them clearer and simpler: [Context] A researcher investigates the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either exposed or not be exposed to the herbicide. The fish exposed to the herbicide showed higher levels of an enzyme associated with cancer. 19. Suppose no statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results? 20. Suppose a statistically significant difference was found between the two groups of fish. What conclusion can be drawn from these results?
20, 21	UCLA Evaluation project (Beckman et al.)	Used the same items that were assessed in a research project [Rob Gould evaluation project] <i>(cont.)</i>

Item Numbers in Preliminary Version AIRS	Item Source and Original Item	Changes Made and Rationale for Change
<i>Table N, cont.</i>		
22	<p>CAOS 37.</p> <p>You have studied statistics and you want to determine the probability of anyone getting at least four right out of six tries just by chance alone. Which of the following would provide an accurate estimate of that probability?</p> <p>a. Have the student repeat this experiment many times and calculate the percentage time she correctly distinguishes between the brands.</p> <p>b. Simulate this on the computer with a 50% chance of guessing the correct soft drink on each try, and calculate the percent of times there are four or more correct guesses out of six trials.</p> <p>c. Repeat this experiment with a very large sample of people and calculate the percentage of people who make four correct guesses out of six tries.</p> <p>d. All of the methods listed above would provide an accurate estimate of the probability.</p>	<p>Modified in wording, questioning and alternatives to emphasize the process of simulating data:</p> <p>a. Specify the probability of a correct guess as 50% and calculate the proportion of all trials with <i>exactly seven</i> correct guesses.</p> <p>b. Specify the probability of a correct guess as 50% and calculate the proportion of all trials with <i>seven or more</i> correct guesses.</p> <p>c. Specify the probability of a correct guess as 70% and calculate the proportion of all trials with <i>exactly seven</i> correct guesses.</p> <p>d. Specify the probability of a correct guess as 70% and calculate the proportion of all trials with <i>seven or more</i> correct guesses.</p>
23-25	<p>Context adapted from CSI project (Allan &amp; Chance) as adapted for use in Robert Gould Evaluation project (Beckman et al.). Items were developed for the topic of Inference about comparing two proportions and Definitions of <i>P</i>-value and statistical significance</p>	
26-31	<p>Adapted from Instructor's Manual and Test Bank for Moore and Notz' (Moore et al., 2008, p. 63)</p>	
32	<p>Created by the author and an Robert delMas</p>	<i>(cont.)</i>

Item Numbers in Preliminary Version AIRS	Item Source and Original Item	Changes Made and Rationale for Change
<i>Table N, cont.</i>		
33-35	Adapted from Instructor's Manual and Test Bank for Moore and Notz' (Moore et al., 2008, p.280)	Topic of Evaluation of statistical testing (considering sample size, practical significance, effect size)
36	<p>CAOS 7.</p> <p>A recent research study randomly divided participants into groups who were given different levels of Vitamin E to take daily. One group received only a placebo pill. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of randomization in this study?</p> <p>a. To increase the accuracy of the research results.</p> <p>b. To ensure that all potential cancer patients had an equal chance of being selected for the study.</p> <p>c. To reduce the amount of sampling error.</p> <p>d. To produce treatment groups with similar characteristics.</p> <p>e. To prevent skewness in the results.</p>	<p>Modified working of the context, questioning, and alternatives to make them clearer and simpler.</p> <p>34. Research participants were randomly assigned to take Vitamin E or a placebo pill. After taking the pills for eight years, it was reported how many developed cancer. Which of the following responses gives the best explanation as to the purpose of randomization in this study?</p> <p>a. To reduce the amount of sampling error that can happen if the subjects are not randomly assigned.</p> <p>b. To ensure that all potential cancer patients had an equal chance of being selected for the study.</p> <p>c. To produce treatment groups with similar characteristics</p> <p>d. To prevent skewness in the results.</p>